Characterization of transcription factors implicated in the regulation of the anticancer indole alkaloid pathway in Catharanthus roseus (L.) G. Don.

Ana Luísa Ferreira Guimarães Silva
Mestrado em Bioquímica
Departamento de Química e Bioquímica

2013

Orientador
Mariana Sottomayor, Auxiliar Professor, Science Faculty, University of Porto
Coorientador
Rosa Cusidó, Full Professor, Pharmacy Faculty, University of Barcelona
Patrícia Duarte, Post-Doctoral Researcher, IBMC

IBMC
INSTITUTO DE BIOLOGIA MOLECULAR E CELULAR
INSTITUTE FOR MOLECULAR AND CELL BIOLOGY

Universitat de Barcelona

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

Agradecimentos

À professora Mariana Sottomayor por, antes de tudo, ter confiado em mim e no meu trabalho para integrar um grupo com a qualidade, o dinamismo e o profissionalismo do Bioactive Natural Products Group. Pelo esforço e pelo empenho que depositou no estabelecimento de uma parceria com o Laboratório de Fisiologia Vegetal da Faculdade de Farmácia da Universidade de Barcelona, proporcionando-me assim uma experiência inesquecível ao abrigo do programa ERASMUS. E, claro, pela orientação durante o último ano. Por tudo o que me ensinou e me deu oportunidade de aprender, por todas as correções e sugestões que seguramente melhoraram este trabalho. Pelo entusiasmo pela ciência e pela investigação que sempre demonstra. Por, a cada novo desafio, me ter ensinado a gostar mais de ciência.

À Professora Rosa Cosidó por tão bem me ter acolhido no Laboratório de Fisiologia Vegetal da Faculdade de Farmácia da Universidade de Barcelona. Por toda a disponibilidade a nível profissional e pessoal. Por todos os conselhos e ensinamentos e pela confiança que depositou em mim. Também ao Professor Javier Palazón pela ajuda e pela boa disposição constantes.

À Patrícia por tudo! Por me ter recebido e se ter disponibilizado para me ensinar o que fosse necessário. Por, com toda a paciência e dedicação, me ter acompanhado diariamente. Por assim, ter tornado este projecto possível e um exemplo de sucesso! Pela preocupação e pelo carinho! Por me transmitir a importância do espírito de grupo e da humildade, mesmo quando já se triunfou. Pela simpatia que nunca deixou de estar presente, mesmo nos piores dias! Um obrigado especial por toda a ajuda e suporte na reta final de escrita da tese. Pelas correções minuciosas, pelas horas e fins de semana que me dedicou, por todos telefonemas, por me ter transmitido a confiança que precisei. Por, também a nível pessoal, sempre se ter disponibilizado para ajudar, por compreender sem julgar. Por ser a amiga e o exemplo que foi neste ano!

A todos os elementos do Bioactive Natural Products Group, por tudo o que me ensinaram e pela companhia diária no laboratório. Um obrigada particular à Sara pela ajuda e disponibilidade, por estar sempre lá! Ao grupo MBNA pelo trabalho em colaboração. À Inês pelos conselhos para a apresentação final deste trabalho.

À Dra. Elisabeth Moyano pela dedicação, pelos conselhos e pela simpatia. A todos os estudantes do Laboratório de Fisiologia Vegetal da Faculdade de Farmácia da

Universidade de Barcelona por toda a ajuda e disponibilidade. Também pela simpatia e por todos os momentos agradáveis que me proporcionaram.

À Cláudia Brito e ao Luís Miguel Santos pela companhia no IBMC.

Ao Ricardo Viais pela partilha dessa grande aventura que foi Barcelona!

Ao Zé por nunca me deixar perder a confiança! Por me encorajar, ouvir e compreender e por acreditar em mim. Por estar sempre presente, ainda que à distância. Por se preocupar comigo, por suportar os meus 'dias maus', por quase nunca perder a paciência. Por antes de mais e acima de tudo ser um grande amigo.

E porque os últimos são os primeiros, aos meus pais! Por me possibilitarem tudo isto. Por me ensinarem a cada dia que a dor do esforço é muito menor do que a dor do fracasso. Por serem sempre os primeiros a acreditar em mim, nas minhas capacidades e no sucesso que vou ter, mesmo quando nem eu acredito! Por estarem sempre disponíveis, por me compreenderem mesmo quando não compreendem. Por respeitarem as minhas escolhas, ainda que por vezes não concordem. Por serem quem são! Por serem o melhor exemplo que algum dia terei!

Resumo

Catharanthus roseus (L.) G. Don acumula nas suas folhas os alcalóides terpenóides indólicos (TIAs) vinblastina (VLB) e vincristina (VCR), que foram os primeiros agentes anticancerígenos de origem natural a serem clinicamente usados, encontrando-se ainda entre os agentes mais valiosos na quimioterapia do cancro. A grande importância farmacológica destes TIAs, aliada à sua baixa abundância na planta (cerca de 0,0005% do peso seco), levou à intensa investigação de *C. roseus*, que se tornou uma das plantas medicinais mais estudadas. No entanto, apesar de se saber muito acerca da biossíntese da via dos TIAs, várias das suas etapas continuam por descrever (tanto a nível genético, como enzimático) e os seus mecanismos de transporte membranar estão completamente por caracterizar. Para além disso, até à data não foi identificado qualquer *master switch* responsável pela regulação desta via. De facto, até hoje nenhuma estratégia para a manipulação da via biossintética dos TIAs em *C. roseus* se revelou eficiente. No entanto, o recurso a fatores de transcrição (FT) com o objetivo de desenvolver estratégias de manipulação para aumentar os níveis de TIAs é tido como uma abordagem promissora, uma vez que os FTs são capazes de regular a transcrição de diferentes genes envolvidos na via de biossíntese dos TIAs.

Em *C. roseus,* os passos tardios limitantes da via biossintética dos TIAs ocorrem em células especializadas de mesófilo chamadas idioblastos, possivelmente o único sítio na planta onde é completada a biossíntese de VLB e VCR. Como tal, os TFs diferencialmente expressos nestas células constituem bons candidatos para a manipulação dos níveis de TIAs.

Neste trabalho foram isolados e clonados com sucesso quatro dos seis FTs em estudo, todos diferencialmente expressos em idioblastos. CrTF12, CrTF19, CrTF79 e CrTF194 foram amplificados com sucesso, após a otimização das condições de PCR, e clonados num vector bacteriano. O CrTF12 foi ainda sub-clonado num vector binário com o promotor 35S e usado para obter *hairy roots* de *C. roseus* que sobreexpressem este FT, num processo mediado por *Agrobacterium.* Este sistema deverá permitir determinar o efeito regulatório da sobreexpressão do CrTF12 na via biossintética dos TIAs. Devido ao crescimento lento das linhas de *hairy roots*, a sua caracterização está ainda em curso. Assim que se obtiver a quantidade necessária de biomassa, os níveis e os perfis dos TIAs, tal como a expressão dos genes da via biossintética, serão analisados.

O CrTF12 foi ainda caracterizado no que respeita à sua localização sub-celular. Para tal, a região codificante deste FT foi clonada na região N-terminal da sGFP e a construção CrTF12-sGFP foi expressa transientemente em protoplastos de mesófilo de *C. roseus,* tendo-se observado uma inequívoca acumulação da proteína de fusão CrTF12-sGFP no núcleo. Este resultado veio reforçar a identidade putativa do CrTF12 como um FT.

O seguimento deste trabalho deverá contribuir para a compreensão da rede transcricional que regula a via biossintética dos TIAs em *C. roseus*, com o objectivo de implementar com sucesso estratégias de manipulação dos níveis dos TIAs anticancerígenos de *C. roseus*, em planta ou em sistemas alternativos de produção *in vitro.*

Palavras-chave: *Catharanthus roseus*, alcaloides indólicos monoterpenoides, fatores de transcrição, *hairy roots*, localização subcelular.

Abstract

Catharanthus roseus (L.) G. Don accumulates in the leaves the dimeric terpenoid indole alkaloids (TIAs) vinblastine (VLB) and vincristine (VCR), which were the first natural anticancer products to be clinically used, and are still among the most valuable agents used in cancer chemotherapy. The great pharmacological importance of the anticancer TIAs, associated with their low abundance in the plant (around 0.0005% DW), stimulated intense research on the TIA pathway, and *C. roseus* has become one of the most studied medicinal plants. However, although much is known about the biosynthesis of TIAs, gene/enzyme characterization is still lacking for many biosynthetic steps, the membrane transport mechanisms of TIAs are basically uncharacterized despite their importance for TIA accumulation, and no effective regulatory master switch of the TIA pathway has been identified to date. In fact, so far, no effective manipulation strategy aiming to increase the TIA levels has ever been successfully achieved. A powerful approach for pathway manipulation is the use of transcription factors (TFs) that are capable of coordinating the transcription of multiple genes involved in a pathway.

In *C. roseus*, the late, bottleneck steps of the TIA pathway take place in specialized mesophyll cells, the idioblasts. These cells accumulate alkaloids and are probably the only place *in planta* where the late TIA pathway occurs. Recently, our laboratory performed a differential transcriptomic profiling of idioblasts and this analysis enabled the identification of six transcription factors (TFs) differentially regulated in the TIA accumulating cells, which are strong candidates for the possible manipulation of TIA levels, and open exciting new avenues of research on TIA metabolism.

In this study, the isolation and cloning of the six TFs differentially expressed in idioblasts was attempted and successfully achieved for four of them. CrTF12, CrTF19, CrTF79 and CrTF194, were successfully amplified after optimization of PCR conditions and cloned in a bacterial vector. CrTF12 was further sub-cloned in a binary vector under the regulation of the strong 35S plant promoter, and used for Agrobacterium mediated transformation of *C. roseus* to generate overexpressing hairy root cultures. This system should now allow to determine the regulatory effect of CrTF12 overexpression in the TIA pathway. The full characterization of the hairy root lines is still underway due to their slow growth. When enough biomass will be obtained, they will be analysed in what concerns the expression levels of TIA related genes and the TIA levels and profile.

CrTF12 was further characterized in what concerns its subcellular localization. A CrTF12-GFP fusion was generated and transiently expressed in *C. roseus* mesophyll protoplasts showing unequivocally the accumulation of the CrTF12-GFP fusion in the nucleus, reinforcing the putative role of CrTF12 as a transcriptional regulator.

The follow up of the present work should provide in the near future an important contribution to the comprehension of the transcriptional network modulating the *C. roseus* TIA pathway, potentially enabling the implementation of successful manipulation strategies to increase the levels of the anticancer TIAs in the plant, or in alternative *in vitro* production systems.

Keywords: *Catharanthus roseus*, terpenoid indole alkaloids, transcription factors, hairy roots, subcellular localization.

Index

List of Figures

List of Tables

Abbreviations

AFLP – Amplified fragment length polymorphism

AS - Anthranilate synthase

AVLB - Anhydrovinblastine

cDNA – Complementary DNA

CiAP - Calf-Intestinal Alkaline Phosphatase

CrTF - *C. roseus* transcription factor

CroPrx1 – *C. roseus* Peroxidase 1

CTAB – Cetyl trimethyl ammonium bromide

D4H - Desacetoxyvindoline 4-hydroxylase

DAT - Deacetylvindoline 4-*O*-acetyltransferase

DMSO - Dimethyl sulfoxide

DNA - Deoxyribonucleic acid

dNTPs - Deoxynucleotide

EDTA - Ethylenediamine tetraacetic acid

EtBr - Ethidium bromide

ER – Endoplasmic reticulum

Fwd - Forward

G10H - Geraniol-10-hydroxylase

10HGO - 10-hydroxygeraniol oxidoreductase

IBA - Indole-3-butyric acid

JERE - Jasmonate and elicitor responsive element

LAMT - loganic acid methyltransferase

MeJA - Methyl jasmonate

MEP - 2-C-methyl-D-erythritol 4-phospate pathway

NLS - Nuclear localization signal

NMT - *N*-methyltransferase

NSCLC - Non-small cell lung cancer

OMT - 16-hydoxytabersonine 16-*O*-methyltransferase

ON - Overnight

ORCA - Octadecanoid-responsive Catharanthus AP2/ERF-domain

ORF – Open reading frame

P450 - Cytochrome P450 monoxygenase

PCR – Polymerase chain reaction

pDNA -Plamid DNA

PEG - Polyethylene glycol

pTi - Tumour-inducing plasmid

pRi- Hairy root-inducing plasmid

RER – Rough endoplasmic reticulum

Rev - Reverse

RNA – Ribonucleic acid

RT – Room temperature

SGD - Strictosidine β-D-glucosidase

sGFP – Green fluorescent protein (S65T)

SLS - Secologanin synthase

STR - Strictosidine synthase

T16H - Tabersonine 16-hydroxylase

Ta - Annealing temperature

TAE buffer – Tris-acetate-EDTA buffer

TDC - Tryptophan decarboxylase

T-DNA - Transferred-DNA

TD-PCR - Touchdown PCR

Tet- Tetracycline

TF - Transcription factors

TGE - Transient gene expression

TIA - Terpenoid Indole Alkaloids

T_L-DNA - Left boarder T-DNA

T_R-DNA - Right boarder T-DNA

UV - Ultra-violet

VCR - Vincristine

VIR - Virulence

VLB - Vinblastine

1 Introduction

1.1 The *Catharanthus roseus* plant

Catharanthus roseus (L.) G. Don. (Figure 1) is a perennial semi-shrub known as the Madagascar periwinkle, which is used as a medicinal plant and that is also commonly appreciated as an ornamental plant. *C. roseus* belongs to the family Apocynaceae and was first published by Linnaeus as *Vinca rosea*, in his "Systema Naturae" in 1759. Later, this species was published as *Catharanthus* by George Don in his "General System of Gardening and Botany" (1835).

Figure 1. Flowering plant of *Catharanthus roseus* cv. Little Bright Eye.

Although *C. roseus* was originally from Madagascar, presently it has a pantropical distribution, being dispersed in Africa, America's, Asia, Australia, Southern Europe, and in some islands in the Pacific Ocean (van der Heijden et al., 2004).

C. roseus has long been used in traditional medicine as an oral hypoglycemic agent in the treatment of diabetes (Singh et al., 2001), and it was the pursuit of this hypoglycemic activity that ultimately led to the discovery of the two anticancer alkaloids present in the leaves of this plant, vinblastine (VLB) and vincristine (VCR) (Noble, 1990; Bennouna et al., 2005). These two complex molecules were classified as alkaloids that possess a cytostatic effect. The high pharmaceutical importance of this anticancer activity stimulated an intense research of *C. roseus*, which has become one of the most studied medicinal plants.

1.2 Alkaloids

Plant secondary compounds can be classified, according especially to their structure but also their biosynthetic origin, in terpenoids, glucosinolates and cyanoglucosides, phenylpropanoids and other phenolic compounds, and finally alkaloids (Giddings,

2011). These compounds are called secondary metabolites because they are not directly involved in the normal growth, development, or reproduction of organisms.

Alkaloids are a group of nitrogen-containing secondary compounds with low molecular weight, usually with basic properties, characteristic for their toxicity and, above all, pharmacological activity. They are mostly derived from the amino acids, Phenylalanine, Tyrosine, Tryptophan, Lysine and Ornithine (De Luca and St Pierre, 2000; Mano, 2006; Memelink and Gantet, 2007). According to their structure, alkaloids can be organized in several groups, and contrasting with other secondary metabolites, there is no biosynthesis or phylogeny correlation between these groups. Research has focused essentially on tropane, pyrrolidine, purine, benzylisoquinoline and the indole alkaloids as a result of their high pharmaceutical interest (Facchini, 2001).

The sessile nature of plants prevents them to look for better environmental conditions, as well as escape from predators. Alkaloids were proved to confer protection from herbivores, pathogens and abiotic stresses, by, for example, avoiding leaf damage from UV light, and their function as allelochemicals has also been proved. All this makes alkaloids and the protection they provide to plants vitally important, although these secondary metabolites are not produced in every condition or developmental stage and do not seem to be essential for the growth or reproduction of plants (Wink, 1993; Bourgaud et al., 2001; Hartmann, 2007). However, it is believed that these compounds confer an adaptable advantage and result from a direct co-evolution of the plant with its surrounding environment (Bourgaud et al., 2001; Sottomayor and Ros Barceló, 2006).

Human societies took advantage of the biological activity presented by alkaloids since early times. In traditional medicine, the latex of opium poppy (*Papaver somniferum*) that contains the alkaloid morphine was already used to treat several illnesses. Moreover, many alkaloids produced by plants are widely used in our everyday life with diverse applications. Such examples are caffeine (from *Coffea arabica* or *Camellia sinensis*) and nicotine (from *Nicotiana tabacum*), used as stimulants, opioid alkaloids (from *Papaver somniferum*), used as sedative or analgesic, and also quinine (from *Cinchona officinalis*), used in the treatment of malaria and lupus and as a bittering agent in tonic water (Mano, 2006; Zhao and Dixon, 2009).

Even though alkaloids with diverse pharmaceutical applications exist, the relevance of the vinca alkaloids (VLB and VCR) in the treatment of cancer, one of the most

deadliest diseases in the wealthiest civilizations, places them among the most important bioactive alkaloids (Sottomayor and Ros Barceló, 2006).

For all the above said and because 25% of all contemporary medicines and 50% of current anti-cancer drugs are derived from plant secondary metabolism, it is clear that these compounds are invaluable sources of newly and unique bioactivities and that their study is of great importance (Memelink and Gantet, 2007). Due to the extreme low levels of most secondary compounds in the plants producing them, one of the main challenges is to find the key regulators of secondary metabolism.

1.3 The terpenoid indole alkaloid pathway in *C. roseus*

Besides the anticancer VLB and VCR (Figure 4), *C. roseus* biosynthesizes a large array of terpenoid indole alkaloids (TIAs) amounting to over 150, including two more with medicinal application, ajmalicine and serpentine, with antihypertensive and sedative properties, respectively (Memelink and Gantet, 2007; Van Moerkercke et al., 2013). TIAs are found in a limited number of plant species belonging mainly to the plant families Apocynaceae, Loganiaceae and Rubiaceae, all from the order Gentianales. Research on the TIA pathway is mainly primed by the pharmaceutical applications of its compounds (Facchini et al., 2004).

TIAs are constituted by an indole moiety and a terpenoid moiety derived respectively from the precursors tryptamine and secologanin (Figure 2) (van der Heijden et al., 2004; O'Connor and Maresh, 2006). Tryptamine results from decarboxylation of the amino acid tryptophan by the action of tryptophan decarboxylase (TDC), which represents the switchover step from primary to secondary metabolism (Loyola-Vargas et al., 2007; Verma et al., 2012). On its turn, tryptophan is originated from the shikimate pathway, where products of the carbohydrate metabolism (phosphoenolpyruvate and erythorose-4-phosphate) are converted into aromatic amino acids, including phenylalanine, tyrosine and tryptophan, which function as precursors of many secondary metabolism pathways (Liu et al., 2007). The monoterpenoid precursor secologanin is derived from geraniol, originated from the 2-C-methyl-D-erythritol 4-phospate pathway (MEP). The conversion of geraniol into secologanin is made via a series of seven putative enzymatic conversions. It is known that one enzymatic conversion is carried out by geraniol-10-hydroxylase (G10H), a cytochrome P450 monoxygenase (P450) enzyme (Zhao and Verpoorte, 2007). The characterization of the remaining steps, hemiacetylation, oxidation, glucosylation, hydroxylation and esterification reactions, is still lacking (Giddings, 2011; Pollier et al., 2011).

The coupling reaction of the indole alkaloid tryptamine with the monoterpenoid secologanin is catalyzed by the enzyme strictosidine synthase (STR), constituting the first committed step of TIA biosynthesis (Figure 2) (Loyola-Vargas et al., 2007; Verma et al., 2012). Strictosidine is the common precursor of all known TIAs and originates distinct pathway branches to generate all the alkaloids produced by *C. roseus* (Memelink et al., 2001).

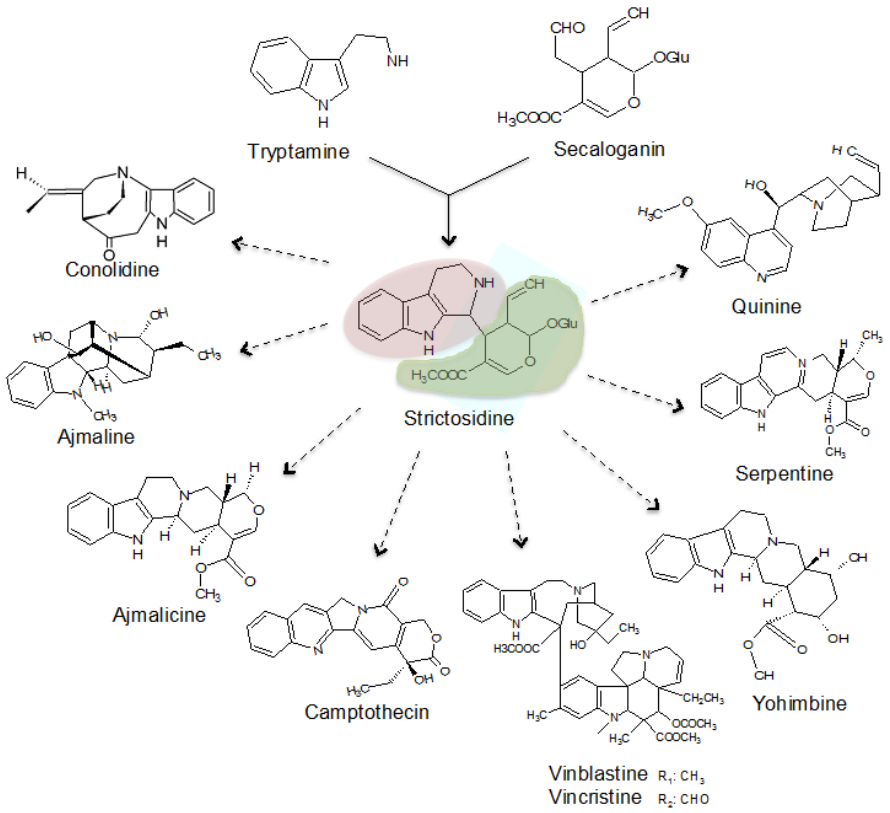


Figure 2. Chemical structure of TIAs with important pharmaceutical activity and of the common precursor of all TIAs, strictosidine. Highlighted in red is the indole moiety of strictosidine, resulting from the indole precursor tryptamine, and highlighted in green is the terpenoid moiety of strictosidine, resulting from the terpenoid precursor secaloganin. Solid arrow's represent single reactions. Dashed arrows represent multiple reactions. *From* Carqueijeiro (2013).

Strictosidine is deglycosylated by strictosidine β-D-glucosidase (SGD) to generate an unstable aglycone that is rapidly converted into a dialdehyde intermediate and further into cathenamine (Figure 3) (Gomez et al., 2009; Guirimand et al., 2011). In *C. roseus*, after cathenamine, the TIA pathway splits in, at least, 3 branches. One of them leads to the production of ajmalicine, which is oxidized to serpentine, in a reaction proposed to be catalyzed by class III peroxidases (Sottomayor et al., 2004). The two other branches

lead to the production of vindoline and catharanthine, the monomeric precursors of the anticancer VLB e VCR (Figure 3).

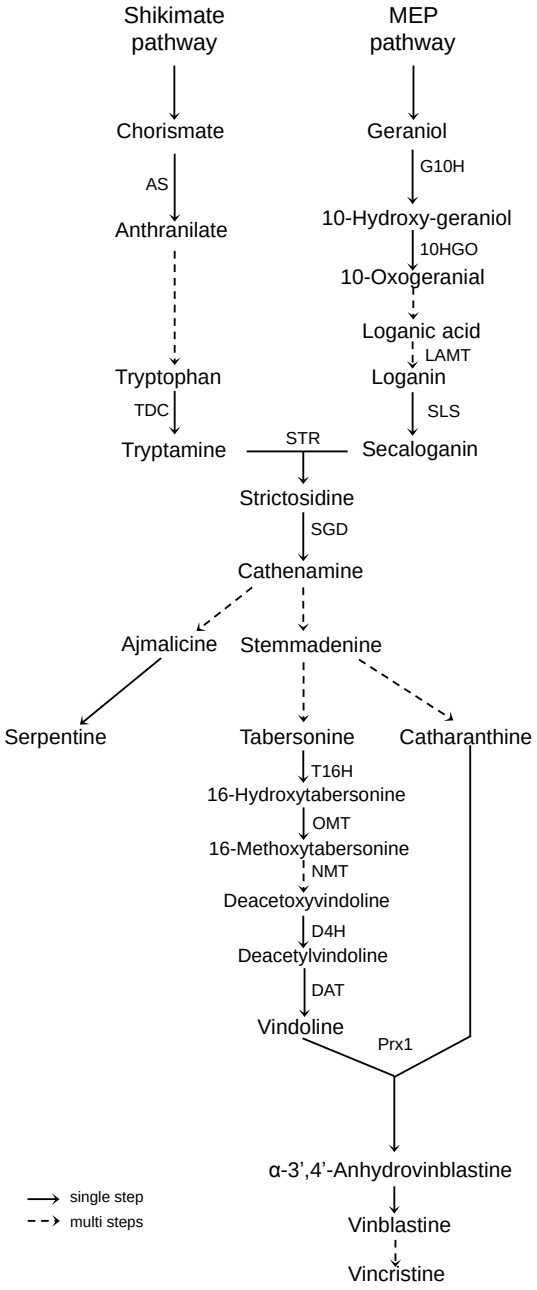


Figure 3. The TIA biosynthetic pathway in *C. roseus*. AS, anthranilate synthase; TDC, tryptophan decarboxylase; G10H, geraniol 10-hydroxilase; CPR, cytochrome P450-reductase; 10HGO, 10-hydroxygeraniol oxidoreductase; LAMT – loganic acid methyltransferase; SLS, secologanin synthase; STR - strictosidine synthase, SGD - strictosidine β-D-glucosidase, T16H - tabersonine 16-hydroxylase, OMT – 16-hydoxytabersonine 16-*O*-methyltransferase, NMT - *N*-methyltransferase, D4H - desacetoxyvindoline 4-hydroxylase, DAT - deacetylvindoline 4-*O*-acetyltransferase, Prx1 – peroxidase 1.

The pathway to catharanthine remains uncharacterized, and the difficulty in the identification not only of the involved enzymes but even of the intermediates, suggests the possibility that most of these reaction intermediates may be unstable, spontaneous and/or highly toxic, and so rapidly converted to the final product catharanthine, which is significantly sequestered in the vacuole (Carqueijeiro et al., 2013).

The last steps of the biosynthesis of vindoline involve six enzymatic reactions starting with the hidroxylation of tabersonine by tabersonine-16-hidroxylase (T16H), a cytochrome P450-dependent enzyme (St-Pierre and De Luca, 1995; Schroder et al., 1999). The resulting 16-hydroxytabersonine is O-methylated by 16-hydroxytabersonine-16-O-methyltranferase (16OMT) (Levac et al., 2008). The third reaction is catalyzed by an unidentified hydratase that performs the hydration of the 2,3-double bond, while the subsequent step of N-methylation involves a partially characterized N-methyltransferase (NMT) producing desacetoxyvindoline (De Luca and Cutler, 1987). The penultimate step, catalyzed by desacetoxyvindoline-4-hydroxylase (D4H), produces desacetylvindoline (De Carolis et al., 1990; Vazquez-Flota et al., 1997). The last step is the acetylation of desacetylvindoline by desacetylvindoline-4-O-acetiltransferase (DAT) to form vindoline (Figure 3) (St-Pierre et al., 1998).

Afterwards, occurs the dimerization of vindoline and catharanthine into the first dimeric alkaloid, α-3',4'-anhydrovinblastine (AVLB), which is the precursor of all dimeric TIAs, including the anticancer VLB and VCR (Figure 3). This dimerization step is believed to be performed by the major class III peroxidase present in the leaves of *C. roseus*, in a reaction characterized in our lab (Sottomayor et al., 1996; Sottomayor et al., 1998; Sottomayor and Ros Barcelo, 2003; Costa et al., 2008). Finally, the biosynthetic steps from AVLB to the anticancer alkaloids, VCR and VLB, are still uncharacterized.

1.4 Bioactivity of the vinca alkaloids

C. roseus has long been used in traditional medicine mainly because of its anti-diabetic properties. It was during the search for novel anti-diabetic drugs in the late 1950s that VLB and VCR, known as the vinca alkaloids, were discovered by Robert Noble and collaborators (Noble, 1990; Singh et al., 2001; Peebles et al., 2011). The usefulness of both VLB and VCR was proved by clinical trials, which led them to be introduced in the clinical shortly after their discovery. The vinca alkaloids became the first natural anticancer agents to be clinically used and they are still an indispensable part of most chemical cocktails used in cancer chemotherapy (Noble, 1990; Sottomayor and Ros Barceló, 2006).

VLB and VCR are both derived from α-3',4'-anhydrovinblastine (AVLB) and are composed of two TIA units differing only in that vincristine has a formyl group at a position where vinblastine has a methyl group, meaning that VCR is the oxidized form of VLB (Figure 4) (Sottomayor and Ros Barceló, 2006; Memelink and Gantet, 2007).

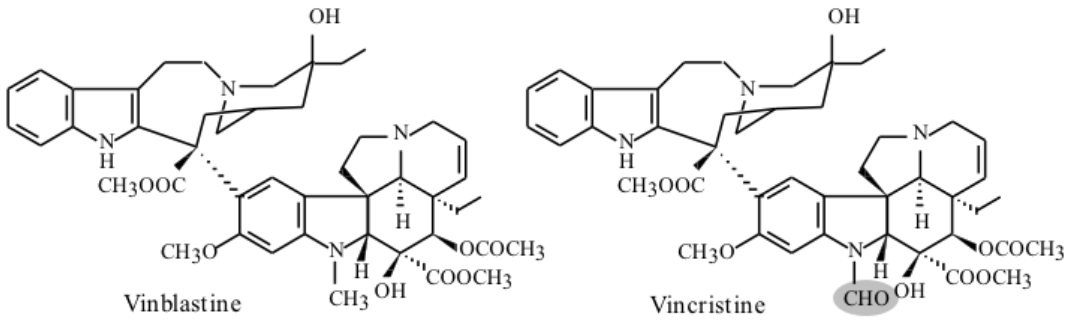


Figure 4. Chemical structure of vinblastine and vincristine. The shaded area indicates the structural difference between the two. *Adapted from* Sottomayor and Ros Barceló (2006).

Despite the minor difference in structure between VLB and VCR, a significant difference exists in the spectrum of human cancers that respond to the drugs, as well as in the toxicity of each one. VLB is mainly used in the treatment of Hodgkin's disease, bladder and breast cancers, and VCR is used in a broader range of cancers, especially in the treatment of leukemia, sarcomas, carcinomas and non-Hodgkin's lymphomas. VCR has superior anti-tumour activity compared to VLB but is also more neurotoxic (Sottomayor and Ros Barceló, 2006; Kingston, 2009; Pollier et al., 2011).

The cytostatic effect of the vinca alkaloids is due to inhibition of cell mitosis as a result from binding to tubulin and inducing tubulin self-association into spiral aggregates. As a consequence, microtubule polymerization is deeply disturbed, impairing mitotic spindle organization in metaphase of the rapidly dividing cancer cells. This inhibits their malignant capacity of division and eventually leads to their death by apoptosis (van der Heijden et al., 2004; Gigant et al., 2005; Loyola-Vargas et al., 2007).

Trying to obtain semi-synthetic derivates of vinca alkaloids, showing higher activity and lower toxicity, has been a high-priority goal for several decades. Successful examples are vinolrebine (Figure 5), launched in 1989, for the treatment of non-metastatic breast cancer and non-small cell lung cancer (NSCLC), with lower neurotoxic side-effects than either VLB or VCR (van der Heijden et al., 2004). Vinflunine (Figure 5), that differs from VLB and vinolrebine in the way it interacts with tubulin, was also tested with success in a broad range of cancers, not only bladder cancer and NSCLC, but also hormone

refractory prostate cancer, ovarian cancer and metastatic breast cancer (Bennouna et al., 2005; Sottomayor and Ros Barceló, 2006). Vindesine (Figure 5) is another semi-synthetic derivative of vinblastine, showing a high activity against lymphoid leukemia in children (Junaid Aslam, 2010).

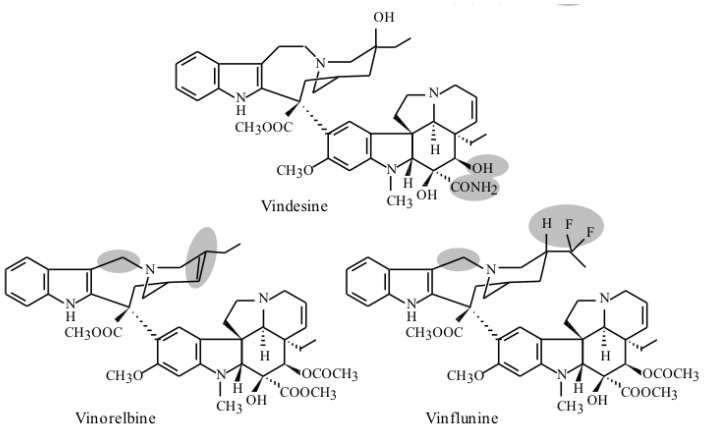


Figure 5. Structure of semi-synthetic vinca alkaloids. Shaded areas indicate the structural differences from vinblastine. *Adapted from* Sottomayor and Ros Barceló (2006).

Although the vinca alkaloids have now earned a place among the most valuable agents used in cancer chemotherapy, their levels in the plant are very low (1 g of VLB per half a ton of dry leaves) (Noble, 1990), what pushed the research of alternative methods for the production of VBL and VCR, namely chemical synthesis and plant cell cultures. However, the chemical synthesis of TIAs is extremely difficult owing to the complexity of their structure, notably the presence of some chiral centers (Zhou et al., 2011). Furthermore, the intricacy of the biosynthetic pathway, involving many diverse steps, makes the chemical synthesis of these compounds currently impossible. Equally, the result of multiple attempts to obtain TIAs in cell cultures was always disappointing, probably because of the very poor alkaloid metabolism that *in vitro* cultures expressed (Hilliou et al., 1999; Sottomayor and Ros Barceló, 2006).

Although VCR and VLB are present in very small amounts in *C. roseus* leaves, their building blocks, vindoline and catharanthine, present higher concentrations, and their use for the semi-synthetic production of the dimers has been investigated. In 1975, Potier and collaborators (1975) reported for the first time the chemical synthesis of AVLB from the coupling of catharanthine and vindoline, with AVLB only later being proved to be the first dimeric biosynthetic precursor of VLB in the plant. Today, although VLB and VCR are still isolated from the plant, vinorelbine and vinflunine are

produced industrially from the monomers by semi-synthetic synthesis (Sottomayor and Ros Barceló, 2006).

1.5 Cellular and sub-cellular compartmentation of the terpenoid indole alkaloid pathway in *C. roseus*

TIAs biosynthesis in *C. roseus* presents organ, tissue, cell and organelle specificity. Catharanthine, ajmalicine and serpentine are the most abundant alkaloids in the roots, whereas vindoline is virtually absent from this organ. Accordingly, D4H and DAT, the genes involved in the last two steps of vindoline biosynthesis, are not expressed in roots. In fact, vindoline is only found in the aerial parts of the plant, especially in the leaves and, likewise, the dimers AVLB, VLB and VCR are only present in these same organs (Facchini and Luca, 2008). Since catharanthine is present in the root in high levels, the question has been raised whether biosynthesis of catharanthine occurs independently in the leaf, or whether it is produced in the root and then transported to the aerial tissues, as observed for instance with nicotine in tobacco (Facchini and Luca, 2008). However, this should not be the case in *C. roseus*, since rootless shoot cultures are capable of producing both catharanthine and vindoline (Hirata K, 1994). Moreover, all TIA intermediates from tabersonine to vindoline were also only detected in the differentiated aerial parts *C. roseus* plants and never in roots (Murata et al., 2008).

In leaves, TIA accumulation has been particularly associated with two specific cell types: unbranched, nonarticulated laticifers associated with the veins, and idioblast cells dispersed along the palisade and spongy mesophyll. Idioblasts are identified by their distinctive blue autofluorescence and by their larger size and lower chloroplast content than the surrounding mesophyll cells (De Luca and St Pierre, 2000; Guirimand et al., 2011; Guirimand et al., 2011). The strong blue fluorescence of idioblasts and also of laticifers has been credited to the TIA serpentine (Brown et al., 1984) (Figure 6) and early experiments using chemical indicators identified laticifers and idioblasts cells as the sites of alkaloid accumulation in *C. roseus* (Yoder, 1976). Mersey and Cutler (1986) have further observed that idioblasts were enriched in vindoline and catharanthine, compared with the remaining mesophyll cells.

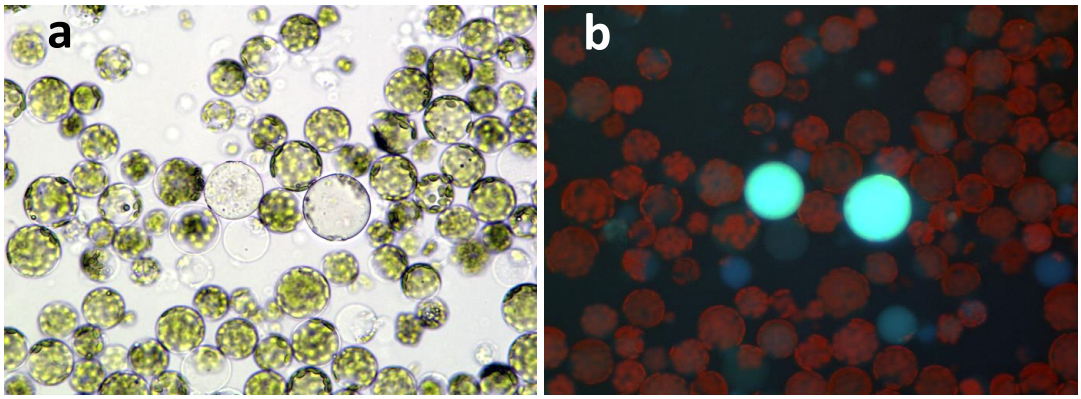


Figure 6. a) Bright field image of *C. roseus* leaf protoplasts. b) Under UV light idioblasts can be distinguished from the rest of the population due to their reduced amount of chloroplasts and conspicuous blue fluorescence because of the presence of serpentine. Red fluorescence corresponds to chlorophyll autofluorescence**.**

More recently, immunocytochemistry and *in situ* RNA hybridization analysis has shown that, in leaves, TIA biosynthesis involves four cell types, including the internal phloem-associated parenchyma (IPAP), the epidermis, laticifers and idioblasts, with each cell-type specifically expressing different genes of the pathway (Figure 7) (St-Pierre et al., 1999; Mahroug et al., 2007; Guirimand et al., 2011; Guirimand et al., 2011). The biosynthesis of secologanin is initiated in the IPAP cells, at least until the hydroxylation of geraniol by G10H (Guirimand et al., 2011). Subsequently, the leaf epidermis is the place where are expressed the early TIA pathway enzymes SLS, TDC, STR, SGD as well as the enzymes catalyzing the first two steps of vindoline biosynthesis. This means that the biosynthesis of the indole precursor tryptamine, of the terpenoid precursor secologanin, and of all TIA intermediates until at least 16-methoxytabersonine take place in leaf epidermal cells (Guirimand et al., 2011; Guirimand et al., 2011). Finally, the expression of DAT and D4H, the enzymes that catalyse the two last steps of vindoline biosynthesis, was specifically detected in leaf idioblasts and laticifers, indicating that these cell types are likely the single place in the plant where the anticancer dimeric VLB and VCR are produced and accumulated (St-Pierre et al., 1999).

Together with the inter-cellular compartmentation, it has been shown that at least four subcellular compartments are also implied in the biosynthesis of TIAs, namely the cytosol, the plastids, the vacuole and the nucleus (Figure 7) (Guirimand et al., 2011; Guirimand et al., 2011; Verma et al., 2012). The shikimate pathway and the biosynthesis of tryptophan take place in the chloroplast, as well as the MEP pathway leading to geraniol (Mahroug et al., 2007). TDC was shown to occur in the cytosol and

G10H was reported to have an ER membrane localization (De Luca and Cutler, 1987; Loyola-Vargas et al., 2007; Guirimand et al., 2009; Guirimand et al., 2011)

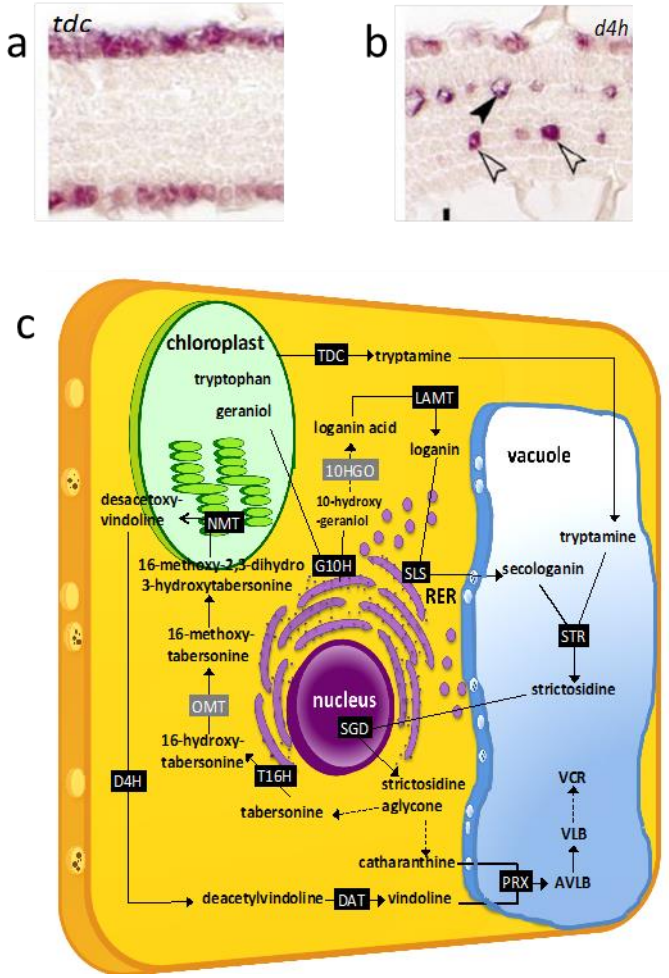


Figure 7. . Celullar (a and b) and subcellular (c) compartmentation of the TIA pathway in *C. roseus* leaves. a and b) Localization of *TDC* (a) and *D4H* (b) RNA in sections of *C. roseus* leaves using *in situ* hybridization; c) Complexity of the subcellular TIA metabolic fluxes in *C. roseus* mesophyll cells. RER - rough endoplasmatic reticulum, G10H - geraniol 10-hydroxylase, SLS - secologanin synthase, TDC - tryptophan decarboxylase, STR- strictosidine synthase, SGD - strictosidine β-D-glucosidade, T16H - tabersonine 16-hydroxylase, OMT - 16- hydoxytabersonine 16-*O*-methyltransferase; NMT - *N*-methyltransferase, D4H - desacetoxy vindoline 4-hydroxylase, DAT - deacetylvindoline 4-*O*-acetyltransferase; PRX– peroxidase. a and b from St- Pierre (1999) and c from Carqueijeiro (2013).

In the cytosol, both LAMT and TDC, form homodimers to prevent their passive diffusion to the nucleus, while SLS is anchored to the cytosolic face of the ER membrane and STR is localized in the vacuole (Mahroug et al., 2007; Guirimand et al., 2010; Guirimand et al., 2011; Carqueijeiro, 2013). SGD was shown to accumulate as highly stable supramolecular aggregates within the nucleus and it was proposed that strictosidine is transported out of the vacuole to suffer deglucosylation in the nucleus (Guirimand et al., 2010). The remaining reactions of the TIA pathway towards tabersonine and catharanthine are poorly characterized, and almost nothing is known

about their subcellular localization. T16H, as expected being a P450, was shown to be anchored to the ER membrane, and OMT was shown to homodimerize in the cytosol (Guirimand et al., 2011). NMT has been proposed to be plastidic, and D4H and DAT were shown to operate as monomers that reside in the nucleocytosolic compartment (De Luca and Cutler, 1987; Guirimand et al., 2011). Finally, CrPrx1 is thought to mediate the coupling of vindoline and catharanthine to yield AVLB, the direct precursor of VLB and VCR, in the vacuole (Costa et al., 2008). The hydroxylation of AVLB into VLB and its subsequent oxidation to VCR may likely also occur in the vacuole. Meaningfully, recent results obtained in our lab showed that the main TIAs accumulated in the leaves, vindoline, catharanthine and AVLB, seem to be exclusively accumulated inside the vacuole (Carqueijeiro, 2013).

The subcellular compartmentation described above implies the existence of multiple steps of transmembrane transport of TIA intermediates, which requires transporter proteins totally uncharacterized until now. Our lab has recently shown that TIA vacuolar accumulation is driven by an H^+-antiport in *C. roseus* mesophyll cells (Carqueijeiro, 2013), suggesting the involvement of a MATE (multidrug and toxic compound extrusion) transporter, since vacuolar import of the alkaloid nicotine in tobacco has been shown to be mediated by MATE transporters functioning as H^+-antiporters (Morita et al., 2009; Shoji et al., 2009).

On top of the subcellular organization of the TIA pathway, the multi-cellular complexity adds further transmembrane transport steps to the TIA pathway, increasing the importance of addressing this aspect of TIA metabolism.

1.6 Regulation of the terpenoid indole alkaloid pathway in *C. roseus*

In planta, the expression of the TIA biosynthetic enzymes appears as a mechanism under tight spatio/temporal molecular regulation, with a complex organ / tissue / cell-specific organization and affected by the development and differentiation stages of the plant, as well as by the environmental conditions (Westekemper P, 1980; Guirimand et al., 2010; Guirimand et al., 2011). This complex spatio/temporal regulation presented by the TIA pathway and other secondary metabolism pathways has been considered a mechanism that plants develop to accomplish the final function of their metabolites.

The majority of the genes codifying TIA biosynthetic enzymes are expressed quite early during development, as both vindoline and catharanthine are accumulated a few days after germination (Vazquez-Flota et al., 1997; De Luca and St Pierre, 2000). In

the developing seedlings, the enzymes TDC and STR are expressed 36-48h prior to D4H and DAT, whose developmental expression, together with vindoline accumulation, is totally dependent on the presence of light and restricted to the aerial parts of the adult pant, mainly in young leaves and cotyledons. It was observed that light treatments activate the late steps of vindoline biosynthesis and also that during leaf development, the concentration of catharanthine and vindoline decreases concomitant with an increase in the levels of their coupling product, AVLB (De Luca and St Pierre, 2000; Costa et al., 2008).

Opposed to D4H and DAT, which are only expressed in the above-ground plant parts, TDC and STR1 occur throughout the plant (De Luca and St Pierre, 2000). TDC is very likely a crucial regulator of the metabolic flux in the TIA pathway, since it is in the intersection between primary and secondary metabolism and therefore the study of this enzyme has been especially extensive (Zhao and Verpoorte, 2007). However, the overexpression of TDC in *C. roseus* hairy roots or crown gall calluses did not result in increasing levels of tryptamine or other TIAs, and the dimeric TIAs were not detected in the roots (Goddijn et al., 1995; Hong et al., 2006).

Expression analysis of the genes encoded TDC and STR1 reveals that both contain sequences involved in the regulation by stress signals (Liu et al., 2011). In agreement, TDC and STR1, have been shown to be often co-regulated, namely when the leaves are exposed to a UV-B light pulse (Ouwerkerk and Memelink, 1999), or when cell suspension cultures are exposed to auxin starvation, fungal elicitors or methyl jasmonate (MeJA) (Goddijn et al., 1992; Pasquali et al., 1992; Menke et al., 1999). Moreover, it has been shown that the induction of TDC and STR expression in response to yeast elicitors depends partially on jasmonates as a secondary signal (Menke et al., 1999). Importantly, MeJA has been considered a regulator of both the early and late TIA pathway, since all the genes tested were induced by MeJA, including D4H and DAT genes (van der Fits and Memelink, 2000; Memelink and Gantet, 2007).

The complex spatio/temporal organization of the TIA pathway and other secondary metabolism pathways is believed to result mainly from transcriptional regulation of the genes encoding biosynthesis enzymes and metabolite transporters, which are expressed in a coordinate manner in response to developmental and environmental signals that activate specific transcription factors (TFs) (Memelink and Gantet, 2007). In *C. roseus*, the tight spatial and temporal regulation of the production of alkaloids suggests the involvement of a strong regulatory network performed by TFs modulating the expression of TIA biosynthetic genes and TIA transporters.

1.7 Plant transcription factors

Transcription factors (TFs) are sequence-specific DNA-binding proteins that are capable of activating and/or repressing transcription. Typical consist of a DNA binding region, an oligomerization site, a transcription regulation domain and a nuclear localization signal (NLS). The well-conserved DNA-binding domains of the TFs had led to the identification of many orphan TFs of unknown function. TFs are largely responsible for the selectivity in gene regulation, and may either be expressed constitutively or in an organ-tissue-specific, developmental-stage-specific, or stimulus-dependent manner (Liu et al., 1999; Zhang, 2003).

Plant TFs are grouped in several families according to their structural similarities. The MYB superfamily is the largest TF family known in plants and it is divided into three families, R2R3, R1R2R3 and MYB-related family (Xiong et al., 2005). MYB genes are very diverse, and only a few of them exhibited constitutive and ubiquitous expression in *Arabidopsis* (Gong et al., 2004; Schmid et al., 2005). MYB proteins are key factors in regulatory networks controlling development, metabolism and responses to biotic and abiotic stresses, and accordingly, they are induced by one or more development signals or by environmental stimuli (Yanhui et al., 2006). Two bHLH TFs have been shown to regulate the anthocyanin pathway in tobacco flowers (Bai et al., 2011).

AP2/EREBP transcriptional regulators are found to be involved in plant key developmental steps, such as flower or seed development and have been implicated in the regulation of the TIA pathway in *C. roseus* (see following chapter). Members of the AP2/EREBP family could be expressed in a tissue-specific manner and/or expressed differentially in response to different phytohormones (Feng et al., 2005). It is believed that this TF family is mainly involved in responses to different stresses (Gutterson and Reuber, 2004).

A conserved WRKY domain that consists of the peptide motif WRKYGQK and a zinc finger is the main characteristic of the WRKY family of TFs (Yamasaki et al., 2005). They form a large, plant-specific TF family and play dynamic roles namely in biotic and abiotic stress responses and salicylic acid signaling (Qu and Zhu, 2006; Rushton et al., 2010). WRKY TFs are known to be involved in alkaloid biosynthesis (Kato et al., 2007; Zhang et al., 2011).

The NAC family constitutes one of the largest families of plant-specific transcription factors, and this family is present in a wide range of land plants. The biologic functions of NAC proteins are, among others, correlated with embryonic, floral and vegetative

development, lateral root formation, auxin signaling and defense. The C-terminal of the NAC proteins is highly diverse and does not contain any known protein domain. On the other hand, the N-terminal presents a conserved region in the encoded protein (Olsen et al., 2005).

Because of TF's pleiotropic action on a wide array of genes involved in metabolic differentiation of plant cells, central TFs enable the development of new strategies to engineer complex metabolic pathways and hold great promise for increasing the levels of pharmaceutically active molecules in plant cells for industrial production (Memelink and Gantet, 2007). A good example of the power of TFs to manipulate a secondary metabolic pathway was the transformation of *Solanum lycopersicum* with two TFs that constitute true master switches of the anthocyanin pathway, inducing the expression not only of many biosynthetic genes, but also of genes codifying anthocyanin transporters (Butelli et al., 2008). It is conceivable that an equivalent architectural organization for the regulation of the TIA pathway in *C. roseus* may exist.

1.8 Transcription factors involved in the regulation of the terpenoid indole alkaloid pathway in *C. roseus*

The TFs ORCA2, ORCA3, CrBPF1, CrWRKY1, CrMYC1 and CrMYC2, and the repressors ZCT1, ZCT2 and ZCT3, are some examples of genes encoding *C. roseus* transcriptional regulators of the TIA pathway which have already been isolated and characterized (Memelink and Gantet, 2007; Suttipanta et al., 2011; Zhang et al., 2011). All of these TFs respond to stress signals, such as jasmonic acid (JA), methyl jasmonate (MeJA), and / or fungal elicitors.

ORCA1 and ORCA2 (octadecanoid-responsive Catharanthus AP2/ERF-domain) are two members of the AP2/EREBP family of TFs and are involved in JA and elicitor responses. Both of them bind to the JERE element (jasmonate and elicitor responsive element, JERE) of the STR promoter, and while ORCA2 responds to MeJA and elicitors, ORCA1 is constitutively expressed an does not respond to those signals (van der Fits et al., 2001; van der Fits and Memelink, 2001; Liu et al., 2011).

ORCA3 is also an AP2/EREBP TF that responds to JA and binds to the JERE element in the STR promoter. Overexpression of ORCA 3 in cell cultures resulted in increasing levels of AS, CPR, SLS, TDC, STR, and D4H (van der Fits and Memelink, 2001; Suttipanta, 2011). However, both G10H and DAT did not respond to ORCA3, suggesting the involvement of other TFs inducible by JA in the regulation of the TIA pathway (Memelink et al., 2001). Transgenic cells overexpressing ORCA3

accumulated significantly more tryptophan and tryptamine, but no TIAs were detected, suggesting that the terpenoid branch of the pathway remained limiting for TIA production. This was confirmed by the fact that when the cells were fed with the terpenoid precursor loganin, ORCA3 overexpression caused an increase in TIA production (van der Fits and Memelink, 2000). Although ORCA3 plays a central role in regulating TIA biosynthesis, it is not sufficient by itself to up-regulate the complete pathway and increase the levels of TIAs. MeJA does not induce TIA biosynthetic gene expression simply by increasing ORCA protein abundance, but instead appears to activate pre-existing ORCA protein. This may occur via post-translational modifications and/or via protein–protein interactions (Vom Endt et al., 2002).

CrBPF1 is a MYB-like transcription factor that also binds to the JERE region of the STR promoter. The accumulation of CrBPF1 is induced by an elicitor but not by MeJA, suggesting that the elicitor induces STR gene expression via JA-dependent and independent pathways (van der Fits et al., 2000). CrMYC1 has been isolated by Chatel et al (2003) and characterized as a basic helix-loop-helix (bHLH). It binds to the STR promoter and may be involved in the regulation of gene expression in response to fungal elicitor and MeJA, since it is induced by these signals.

C. roseus ZCT1, 2 and 3 proteins are members of the Cys2/His2-type zinc finger gene family. These proteins repress the activity of the TDC and STR promoters in trans-activation assays, by binding them in a sequence-specific manner *in vitro*. Curiously, the expression of ZCTs was induced by MeJA and yeast extracts, similarly to what happens with the ORCA activators. This paradox of the induction of both activators and repressors of STR and TDC gene expression by elicitor and MeJA was suggested by the authors to serve to fine tune the amplitude and timing of gene expression (Pauw et al., 2004).

Another bHLH TF, CrMYC2, was shown to be involved in the MeJA induction of TIA biosynthetic genes, since the decrease of the CrMYC2 expression level caused a strong reduction in the level of MeJA-responsive ORCA3 and ORCA2 mRNA accumulation (Zhang et al., 2011). Overall, the results showed that the MeJA-responsive expression of TIA biosynthesis genes in *C. roseus* is controlled by a TF cascade consisting of CrMYC2 regulating the expression of ORCA genes, which, in turn, regulate a subset of TIA biosynthetic genes.

CrWRKY1 is a *C. roseus* WRKY TF described by Suttipanta et al (2011), that binds to the W box elements of the TDC promoter. It is preferentially expressed in roots and

induced by the phytohormones jasmonate, gibberellic acid, and ethylene. The overexpression of CrWRKY1 in *C. roseus* hairy roots leads to the up-regulation of TDC, ZCT1, ZCT2, and ZCT3, and to the repression of the transcriptional activators ORCA2, ORCA3, and CrMYC2. Compared with control roots, CrWRKY1 hairy roots accumulated up to 3-fold higher levels of serpentine. The preferential expression of CrWRKY1 in roots and its interaction with transcription factors including ORCA3, CrMYC2, and ZCTs may play a key role in determining the root-specific accumulation of serpentine in *C. roseus* plants.

Although regulators of several steps of the TIA biosynthetic pathway have been characterized, so far, it was not possible to isolate TFs that may function as master switches leading to an increase of the levels of the anticancer TIAs. This is possibly because idioblast-specific TFs regulating the late, bottleneck part of the pathway were never identified since most of these studies were performed with cell cultures that do not express the late part of the TIA pathway (Memelink and Gantet, 2007; Zhang et al., 2011).

In a work performed in our lab, the leaf TIA accumulating idioblast cells were isolated by fluorescence activated sorting of mesophyll protoplasts, and their differential transcriptomic profiling by cDNA-AFLP enabled the identification of a number of candidate genes potentially involved in the late TIA pathway. Among the genes found to be up-regulated in idioblasts several TFs from different families were identified (Carqueijeiro, 2013). Since the major regulatory mechanism modulating alkaloid production in plant cells is thought to be the transcriptional control of the biosynthetic genes, it is expectable that the identified CrTFs may include important master switches of the late, bottleneck TIA pathway. These CrTFs may therefore be used, in the future, to transform *C. roseus* cells and plants to generate overexpression lines potentially producing higher levels of TIAs.

Table 1. List and main features of the TFs studied in this work. -:without expression; †: low expression; ††: medium expression; †††: high expression.

Short name	Annotation	Roots	Leaves	Protoplasts	Idioblasts
CrTF12	NAC domain-containing protein 867 bp	-	-	†	††
CrTF19	DNA binding protein 1920 bp	†	†	†	†
CrTF79	No match 1452 bp	-	-	††	-
CrTF152	*A. thaliana* TRFL2 (TRF-LIKE 2); DNA binding (TRFL2)/A. thaliana MYB TF 2090 bp	†	†	-	††
CrTF194	Myc2 bHLH protein [*Vitis vinifera*] 1812 bp	†	†	†	†††
CrTF246	Homeobox protein 2184 bp	-	-	-	†

1.9 Objectives

The objective of this project was to perform the isolation and cloning of transcription factors (TFs) differentially expressed in *C. roseus* TIA accumulating cells, the idioblasts, since those novel CrTFs are strong candidates to perform an important role in the regulation of the anticancer TIA pathway and/or idioblast differentiation.

Four CrTFs were successfully cloned, after which the objectives were to proceed for functional characterization of a first CrTF including: i) overexpression in transgenic *C. roseus* hairy root cultures[1] followed by evaluation of its impact on the expression level of TIA biosynthetic genes by qPCR and on the TIA levels/profile by HPLC, and ii) characterization of subcellular localization by generation of a CrTF-GFP gene used for transient expression in *C. roseus* mesophyll protoplasts. Hairy roots overexpressing CrTF12 were generated and are in the process of producing enough biomass to perform the subsequent experiments. Transient expression of a CrTF12-GFP fusion in *C. roseus* protoplasts indicated localization in the nucleus, reinforcing the putative role of CrTF12 as a transcriptional regulator.

We hope that this work may provide a contribution to the comprehension of the transcriptional network modulating the *C. roseus* anticancer TIA pathway, with the future goal of implementing successful manipulation strategies to increase the levels of the anticancer TIAs in the plant, or in alternative *in vitro* production systems.

[1] See Appendix 1 for an introduction to Agrobacterium mediated transformation and hairy root cultures.

2 Materials and Methods

2.1 Biological Material

2.1.1 Plant Material

Catharanthus roseus (L.) G. Don cv. Little Bright Eye seeds were acquired from B&T World Seeds (France) and germinated in SiroPlant compost (Siro). The pots were initially covered with cling film to start the acclimatization process and then were regularly transferred into bigger pots as growth proceeded. Plants were maintained in a growth chamber at 25°C, under a 16 h photoperiod, using white fluorescent light with a maximum intensity of 70 µmol m^{-2} s^{-1}. For the isolation of protoplasts, the 2nd and 3rd leaf pairs of adult *C. roseus* plants were used.

2.1.2 Bacterial strains and plasmids

In this work, the strains *Escherichia coli* DH5α and TOP10, *Agrobacterium rhizogenes* A4, as well as *Agrobacterium tumefaciens* C58C1 were used.

Chemically competent *E. coli* cells were prepared according to a protocol adapted from Hanahan et al. (1991), described in Appendices 2 and 3. *Agrobacterium* competent cells were also chemically obtained, following the protocol described in Appendix 4.

For the primary cloning of the full coding sequence of the transcription factors under study, the commercial vector pGEM®-T easy (Promega) was used.

To perform the subcellular localization of CrTF12 by transient expression in *C. roseus* mesophyll protoplasts, a fusion of the CrTF12 coding sequence was generated at the 5'-end (N-terminus) of sGFP using the plasmid pTH2. pTH2 corresponds to pUC18 carrying the 35S Ω-sGFP(S65T)-nos construct and an ampicillin/carbenicillin-resistance marker (Niwa et al., 1999).

The main features of the plasmids used along this work are shown in Table 2.

Table 2. List and main features of the plasmids used in this work.

Vector	Type	Size (bp)	Promoter	Type of promoter
pGEM-Teasy	Cloning vector	3015	-	-
pTH2	Plant Vector	4090	35S	Constitutive
pGreenII35S	Binary Plant Expression Vector	8299	35S	Constitutive

2.2 *In silico* analysis of sequences

The tags for CrTFs differentially expressed in *C. roseus* idioblasts (specialized TIA accumulating cells) were obtained using a differential transcriptomic approach (cDNA-AFLP) performed with roots, leaves, protoplasts and idioblasts (Carqueijeiro, 2013).

The full-length cDNA sequences of CrTF12, CrTF19, CrTF79, CrTF152, CrTF194, CrTF246 were obtained using the transcriptomic databases MPGR (http://medicinalplantgenomics.msu.edu/index.shtml) and PhytoMetaSyn (http://www.phytometasyn.ca/). Identification of the main features of each CrTF family was investigated using the Plant Transcription Factor Database (http://plntfdb.bio.uni-potsdam.de/v3.0/). The translation of the open reading frames of the obtained nucleotide sequences was performed using the Translate tool of ExPASy Bioinformatics Resource Portal (http://web.expasy.org/translate/). Finally, *in silico* subcellular localization was predicted using TargetP 1.1 (http://www.cbs.dtu.dk/services/TargetP/) (Emanuelsson et al., 2000) and WoLF PSORT (http://www.genscript.com/psort/wolf_psort.html) (Horton et al., 2007).

2.3 Isolation of full length coding-sequences of the transcription factors

2.3.1 Extraction of RNA and synthesis of cDNA

Total RNA was extracted from leaves at different stages of development (immature and fully expanded) from 4 month-old and mature plants (> 7 months), and also from leaf mesophyll protoplasts from plants of the above mentioned ages. Seedlings (*ca.* 1 month) were also used as starting material for RNA extraction. RNA was obtained using the RNeasy Plant Mini kit (Qiagen), according to the manufacturer's instructions.

Each sample contained a pool of biological material composed of leaves/protoplasts from 2 to 3 plants. The extracted RNA was quantified spectrophotometrically in a Nanodrop ND-1000 (Thermo Scientific). For the synthesis of cDNA, 1 µg of the isolated RNA was first treated with DNaseI (Thermo Scientific) to eliminate DNA contamination in further reactions. The enzyme was then inactivated by heating at 65°C for 10 min in the presence of 20 mM EDTA. First-strand cDNA was synthesized from total RNA using the iScript™ cDNA Synthesis Kit (BioRad) according to manufacturer's instructions, with oligo(dT) as primary reverse primers for the reaction.

2.3.2 Design of primers

Primers for primary cloning into pGEM® T-easy (Promega) were designed to amplify the full coding sequence of the transcripts and perform TA cloning. For the cloning into the binary vector, pGreenII35S and into the expression vector pTH2, the primers included endonuclease restriction sites to allow directional cloning. The reverse primer used for the cloning into the expression vector pTH2, also had a linker sequence shown to help in fusion protein folding, in this case of, both CrTF12 and sGFP (encoded by pTH2). To ensure that the endonuclease restriction sites included in the primers were not present in the coding sequence of the CrTFs, the NEBcutter V2.0 tool was used (http://tools.neb.com/NEBcutter2/). All relevant parameters of the primers such as melting temperature (Tm), stability of secondary structures, and propensity to form homo- and hetero-dimers were evaluated using the OligoAnalyzer tool of Integrated DNA Technologies, (http://eu.idtdna.com/analyzer/applications/oligoanalyzer/default.aspx) (Table 3).

Table 3. Primers used to amplify the full length coding sequences of the TFs for cloning in different vectors.

Primer	Sequence (5' to 3')	Size (bp)	Tm (°C)
Primary Cloning			
CrTF12 Fwd	ATGGCTGCAGAGTTGCAATTACC	23	58.2
CrTF12 Rev	TTAAAATGGCTTCTGCAAGAACATGAAC	28	56.7
CrTF19 Fwd	ATGGAAGGAACAATTTGTTTCGAAGG	26	56.2
CrTF19 Rev	TTAACATTCCTGCCTGGCTG	20	55.4
CrTF79 Fwd	ATGAAGGTTACGGAGAAGGATATGTC	26	56.0
CrTF79 Rev	CTAATATTGTGAAACCCAATCATAAAGGGC	30	56.4
CrTF152 Fwd	ATGGTGTTGCAGAACAGGTTAG	22	55.3
CrTF152 Rev	TTTCAAAGAAGGCGGCAAGC	20	56.8
CrTF194 Fwd	ATGAAAATAGGAGTCGGTATGGGAGG	26	57.9
CrTF194 Rev	TCACGAAGATTGTGAGGTTGCC	22	57.6
CrTF246 Fwd	ATGTTTGATAGTCACCATCACTTAC	25	52.9
CrTF246 Rev	CTAAGCATTTTCACAGGCTAATG	23	52.2
Sub-cloning into pGreenII35S			
CrTF12 pGreen Fwd	ATGGCGCGCCATGGCTGCAGAGTTGCAATTAC	32	68.8
CrTF12 pGreen Rev	ATGCGGCCGCTTAAAATGGCTTCTGCAAGAAC	32	65.8
CrTF19 pGreen Fwd	ATGGCGCGCCATGGAAGGAACAATTTGTTTCG	32	65.9
CrTF19 pGreen Rev	ATGCGGCCGCTTAACATTCCTGCCTGGCTG	30	68.9
CrTF79 pGreen Fwd	ATGGCGCGCCATGAAGGTTACGGAGAAGG	29	67.1
CrTF79 pGreen Rev	ATGCGGCCGCCTAATATTGTGAAACCCAATC	31	63.9
CrTF152 pGreen Fwd	ATGGCGCGCCATGGTGTTGCAGAACAGGTTAG	32	68.1
CrTF152 pGreen Rev	ATGCGGCCGCTTTCAAAGAAGGCGGCAAG	29	68.3

CrTF194 pGreen Fwd	ATGGCGCGCCATGAAAATAGGAGTCGGTATG	31	64.9
CrTF194 pGreen Rev	ATGCGGCCGCTCACGAAGATTGTGAGG	27	66.3
CrTF246 pGreen Fwd	ATGGCGCGCCATGTTTGATAGTCACCATCAC	31	65.3
CrTF246 pGreen Rev	ATGCGGCCGCCTAAGCATTTTCACAGGCTAATG	33	66.5
Sub-cloning into pTH2			
CrTF12 pTH2 N-terminal Fwd	ATGGCGCGCCATGGCTGCAGAGTTGCAATTAC	32	68.8
CrTF12 pTH2 N-terminal Rev	TCTGGTTCAGGATCCATGG	19	54.0

2.3.3 Polymerase Chain Reaction (PCR)

Firstly, all primer combinations were tested performing PCR with *DreamTaq* DNA polymerase (Thermo Scientific) to confirm primer specificity and the expected size of the amplification products. All the variables assayed in the amplification optimization process are described in Table 4.

Table 4. Variables assayed for the optimization of the initial PCR reactions, for the amplification of each TF.

Parameter	Tested conditions
DNA polymerase	*DreamTaq* (Thermo Scientific); *FideliTaq* ® (Affimetrix)
Template (µL)	1; 2
Ta (°C)	48 – 54
PCR enhancers	1M Betaine; 5% DMSO; 1M Betaine + 5% DMSO
Number of cycles	35; 40; 50
Extension temperature (°C)	68; 72
PCR type	Conventional PCR; Touchdown PCR

For cloning purposes in pGEM® T-easy (Promega), the composition of the PCR reactions was the typical, shown in Table 5, and these reactions were performed with *DreamTaq* DNA polymerase (Thermo Scientific) and *DreamTaq* DNA polymerase buffer including 20 mM $MgCl_2$. 1M of Betaine and 1M Betaine + 5% DMSO were used for amplification of CrTF194 and CrTF152, respectively.

Table 5. Typical PCR reaction, using *DreamTaq* DNA polymerase.

PCR reagents	Volumes
***DreamTaq* DNA polymerase**	2.5
cDNA template	2
10 mM dNTPs (Thermo Scientific)	0.5
10 µM Primer Fwd	1
10 µM Primer Rev	1
***DreamTaq* DNA polymerase (Thermo Scientific)**	0.2
Sterile H_2O	17.8
Final volume	25

The thermal cycling conditions are shown in Table 6. For CrTF246, in order to obtain a better yield, two rounds of PCR were performed.

Table 6. PCR conditions, using *Taq* DNA polymerase, for the amplification of each TF.

CrTF12						
	95°C	**95°C**	**48°C**	**72°C**	**72°C**	**Hold 6°C**
Time	3 min	30 sec	45 sec	2.5 min	7 min	∞
Number of cycles	1	40			1	
CrTF19						
	95°C	**95°C**	**50°C**	**72°C**	**72°C**	**Hold 6°C**
Time	3 min	30 sec	45 sec	2.5 min	7 min	∞
Number of cycles	1	40			1	
CrTF79 and CrTF194						
	95°C	**95°C**	**49°C**	**72°C**	**72°C**	**Hold 6°C**
Time	3 min	30 sec	45 sec	2.5 min	7 min	∞
Number of cycles	1	40			1	
CrTF152						
	95°C	**95°C**	**50°C**	**72°C**	**72°C**	**Hold 6°C**
Time	3 min	30 sec	45 sec	2.5 min	7 min	∞
Number of cycles	1	40				1
CrTF246						
1st PCR round	**95°C**	**95°C**	**46°C**	**72°C**	**72°C**	**Hold 6°C**
Time	3 min	30 sec	45 sec	3 min	7 min	∞
Number of cycles	1	50			1	
2nd PCR round	**95°C**	**95°C**	**49°C**	**72°C**	**72°C**	**Hold 6°C**
Time	3 min	30 sec	45 sec	3 min	7 min	∞
Number of cycles	1	40			1	

After primary cloning of the CrTFs that we were able to amplify, all clones were sequenced and were checked to be error-free.

A DNA polymerase with proofreading activity (*Pfu*; Thermo Scientific) was used for the amplification of the products to be cloned in pGreenII35S. The primer combinations for CrTF12 and CrTF79 were tested to confirm primer specificity and the expected size of the amplification products. The variables assayed on the optimization process are described in Table 7. Tables 8 and 9 describe, respectively, the composition of the PCR reactions and the thermal cycling conditions used in these reactions

Table 7. Variables assayed for the optimization of the PCR reactions, using *Pfu* DNA polymerase, for the amplification of CrTF12 and CrTF79.

	95°C	**95°C**	**50-55°C**	**72°C**	**72°C**	**Hold 6°C**
Time	3 min	30 sec	45 sec	4 min	10 min	∞
Number of cycles	1	40			1	

Table 8. PCR reaction, using *Pfu* DNA polymerase, for the amplification of CrTF12 and CrTF79.

PCR reagents	**Volume (µL)**
10X *Pfu* Buffer with $MgSO_4$	5
pDNA (1:10)	2
10 mM dNTPs (Thermo Scientific)	2
10 µM Primer Fwd	2
10 µM Primer Rev	2
***Pfu* DNA polymerase (Thermo Scientific)**	0.5
Sterile H_2O	36.3
Final volume	50

Table 9. PCR conditions, using *Pfu* DNA polymerase, for the amplification of CrTF12 and CrTF79.

CrTF12						
	95°C	**95°C**	**55°C**	**72°C**	**72°C**	**Hold 6°C**
Time	3 min	30 sec	45 sec	4 min	10 min	∞
Number of cycles	1	40			1	
CrTF79						
	95°C	**95°C**	**50°C**	**72°C**	**72°C**	**Hold 6°C**
Time	3 min	30 sec	45 sec	4 min	10 min	∞
Number of cycles	1	40			1	

In order to clone CrTF12 in the expression vector pTH2, the following annealing temperatures were tested: 52°C, 54°C, 56°C, 58ºC and 60°C. The composition of the PCR reaction was the typical (Table 5) and the PCR program is shown in Table 10.

Table 10. PCR conditions, using *DreamTaq* DNA polymerase, for amplification of CrTF12.

	95°C	**95°C**	**56°C**	**72°C**	**72°C**	**Hold 6°C**
Time	3 min	30 sec	45 sec	4 min	7 min	∞
Number of cycles	1	40			1	

2.3.4 Touchdown PCR

In order to obtain a higher yield in the primary amplification of CrTF246, touchdown PCR (TD-PCR) was performed. The programme used for TD-PCR reactions is shown in Table 11.

Table 11. TD-PCR conditions, using *DreamTaq* DNA polymerase, for the amplification of CrTF246.

18 first cycles					
	95°C	**95°C**	**55°C (-0.5°C/cycle)**	**72°C**	
Time	3 min	30 sec	45 sec	2.5 min	
Number of cycles	1	18			
Remaining 22 cycles					
	95°C	**46°C**	**72°C**	**72°C**	**Hold 6°C**
Time	30 sec	45 sec	2.5 min	10 min	∞
Number of cycles	22			1	

The reaction was performed with *DreamTaq* and different enhancers and enhancer combinations were tested (Table 12).

Table 12. TD-PCR reaction, using *DreamTaq* DNA polymerase, for the amplification of CrTF246.

PCR reagents	**Without enhancers**	**5M Betaine**	**5% DMSO**	**1M Betaine + 5% DMSO**
10X *DreamTaq* Buffer	2.5	2.5	2.5	2.5
cDNA template	2	2	2	2
10 mM dNTPs (Thermo Scientific)	0.5	0.5	0.5	0.5
10 µM Primer Fwd	1	1	1	1
10 µM Primer Rev	1	1	1	1
***DreamTaq* DNA polymerase (Thermo Scientific)**	0.2	0.2	0.2	0.2
1M Betaine	-	5	-	5
5% DMSO	-	-	1.25	1.25
Sterile H_2O	17.8	12.8	16.55	11.55
Final volume	25	25	25	25

2.3.5 Visualization and Purification of PCR products

The PCR products were visualized in 1% (w/v) agarose (Bio-rad) gel electrophoresis using Tris-acetate-EDTA [TAE buffer; 40 mM Tris base (Sigma), 10% (v/v) acetic acid (Merck) and 10 mM EDTA (Merck)] as running buffer. The GeneRuler™ 1kb DNA Ladder (0.5 µg μL^{-1}; Thermo Scientific) was used as molecular marker and an Orange Loading Dye Solution (Sigma) was used for visual tracking of DNA migration. Ethidium bromide (EtBr; 10 mg mL^{-1}; BioRad) was added to the agarose gel at a 0.5 µg mL^{-1} concentration, to allow visualization of the DNA bands under UV light. The electrophoretic run was performed at 80-100 V (PowerBac Basic, BioRad).

The DNA of interest was recovered from the agarose gel using the GeneJET™ Gel extraction kit (Thermo Scientific), according to the manufacturer's instructions.

2.4 Molecular cloning

2.4.1 Ligation to cloning vectors

All full coding sequences amplified without restriction sites were cloned into pGEM®-T Easy (Promega), according to the manufacturer's instructions for TA cloning.

In order to clone the CrTF12 and CrTF79 in the binary vector pGreenII35S, both the amplification products of CrTF12 and CrTF79 (obtained with the pGreen primers - Table 3) and the pGreenII35S vector were double digested with *Not*I and *Sgs*I (*Asc*I) (all from Thermo Scientific) restriction enzymes, following the manufacturer's instructions. The digestion reactions were composed as described in Table 13 and were incubated at 37ºC during 3 h. To avoid pGreenII35S plasmid re-ligation, 1 µL of Calf-Intestinal Alkaline Phosphatase (CiAP, Thermo Scientific) was added to the reaction in the last hour of incubation.

After this reaction time, restriction enzymes were inactivated by adding 20 mM EDTA (pH 8.0) and further incubating the mixture at 80ºC for 20 min.

Table 13. Composition of the restriction reactions performed to the binary vector pGreenII35S and the PCR purified inserts CrTF12 and CrTF79.

Restriction Reagents	**Volume (µL)**	
	Purified PCR products	**pGreenII35S**
DNA	20	10
Buffer R (1x)	8	8
***Not* I**	4.5	4.5
***Sgs*I (*Asc*I)**	3.4	3.5
Sterile H_2O	44	54
Final Volume	80	80

The digested DNA was recovered from an agarose gel and its integrity was tested as described above. Ligation reactions were set with T4 DNA ligase (Thermo Scientific) as recommended by the manufacturer (Table 14), and incubated at RT for 3 h.

Table 14. Composition of the ligation reactions performed to clone the CrTF12 and CrTF79 in pGreenII35S.

Ligation Reagents	**Volume (µL)**
pGreenII35S	1
CrTFs	3
10x T4 DNA Ligase Buffer	1
T4 DNA Ligase	1
Sterile H_2O	4
Final Volume	10

With the purpose of sub-cloning the CrTF12 in the expression vector pTH2, both the pTH2 vector and the amplification products of pGEM®-T Easy-Cr12 with CrTF12 pTH2 N-terminal primer (Table 3) were double digested with *Sal*I and *Nco*I (all from Thermo Scientific) restriction enzymes, following the manufacturer's instructions (Table 15).

Table 15. Composition of the restriction reactions performed to the expression vector pTH2 and the PCR purified inserts CrTF12.

Restriction Reagents	**Volume (µL)**	
	Purified PCR products	**pTH2**
DNA	20	10
Buffer Tango (2x)	16	16
***Sal*I**	4.5	4.5
***Nco*I**	3.5	3.5
Sterile H_2O	41	41
Final Volume	80	80

The cut DNA was recovered and checked as described before. Ligation reactions (Table 16) were incubated at RT for 3h.

Table 16. Composition of the ligation reactions performed to clone the CrTF12 in pTH2.

Ligation Reagents	**Volume (µL)**
pTH2	1
CrTF12	3
10x T4 DNA Ligase Buffer	1
T4 DNA Ligase	1
Sterile H_2O	4
Final Volume	10

2.4.2 Transformation of *E. coli* and selection of positive colonies

All ligation products were transformed into *E. coli* DH5α or *E. coli* TOP10 using the heat shock method. Typically, 50-100 µL of *E. coli* chemically competent cells were used for transformation either with 1-2 µL of plasmid DNA (pDNA) or half of the volume of a ligation reaction (5 µL). After an incubation of 30 minon ice, the mixture of competent bacteria and DNA was heat-shocked at 42°C for 2 min and placed back on ice for 2 min. One mL of LB (Appendix 2) was added to the mixture and the cells were left to recover for 1h at 37°C. The cell suspension was centrifuged at 1500 g (4000 rpm) for 3 min at room temperature (RT) and 900 µL of supernatant were removed. The cells were resuspended in the remainder volume, plated onto LB-agar (LB with 1.5% agar;

Liofilchem) supplemented with 100 µg mL^{-1} of ampicillin (for pGEM®-T Easy and pTH2 vectors) or 25 µg mL^{-1} of kanamycin (for pGreenII35S vector), and the plates were incubated ON at 37°C. At least 3 colonies of each ligation product were selected to grow in 5 mL of LB medium supplemented with the appropriate antibiotic and were incubated ON at 37°C with vigorous shaking (200 rpm). These liquid cultures were used to check for the presence of the correct pDNA using the Plasmid Miniprep Kit, Thermo Scientific according to the manufacturer's instructions.

Cloning of TFs into vectors, pGEM®-T Easy (Promega), pGreenII35S and pTH2, was confirmed by restriction analysis of the miniprep DNA, performed at 37°C (Table 17, Table 18 and Table 19).

Table 17. Composition of the digestion reaction performed to select the positive clones, transformed with pGEM®-T Easy-TF.

Restriction Reagents	**Volumes (µL)**
pGEM®-T Easy-TF (mp)	2
Buffer O (1x)	2
***Not*I**	1
Sterile H_2O	15
Final Volume	20

Table 18. Composition of the digestion reaction performed to select the positive clones, transformed pGreenII35S-CrTF.

1st Step	
1st step reaction reagents	**Volumes (µL)**
pGreenII35S-TF	6
Buffer Tango (1x)	5
***Sgs*I (*Asc*I)**	1
Sterile H_2O	38
Final Volume	50
2nd step	
Reagents added	**Volumes (µL)**
Buffer Tango (2x)	8
***Not*I**	4
Sterile H_2O	3
Final Volume	65

Table 19. Composition of the digestion reaction performed to select the positive clones, transformed with pTH2 – CrTF12.

Restriction Reagents	Volumes (µL)
pTH2 – CrTF12(mp)	2
Buffer O (1x)	2
***Sal* I**	0.5
Sterile H_2O	15.5
Final Volume	20

The inserted nucleotide sequences were confirmed by sequencing (STAB Vida) and analysis of sequence homology between the predicted sequences and the obtained sequences was performed using the multiple sequence alignment program MultiAlin (http://multalin.toulouse.inra.fr/multalin/).

2.4.3 Transformation of *Agrobacterium*

To co-transform *Agrobacterium* chemically competent cells with both pGreenII35S, (empty or harbouring the CrTF coding sequence), and the helper plasmid pSoup, 1µL of each pDNA (Mini-prep) was added to 100 µL of cells. The mixture of competent *Agrobacterium* cells and DNA was incubated 5 min on ice, then placed in liquid nitrogen for another 5 min and finally, was heat-shocked at 37°C for 5 min. One mL of YEB (Appendix 4) was added to the mixture and the cells were left to recover for 3h at 28°C. The cell suspension was centrifuged at 1500 g (4000 rpm) for 3 min at RT and 900 µL of supernatant were removed. The cells were resuspended in the remainder volume, were plated onto YEB-agar (YEB with 1.5% agar; Liofilchem), supplemented with 100 µg mL^{-1} of rifampicin, 25 µg mL^{-1} of kanamycin and 5 µg mL^{-1} of Tetracycline, and the plates were incubated two ON at 28°C. The grown bacteria were used to infect *C. roseus* leaves.

2.5 *Agrobacterium* Infection and hairy root generation

C. roseus leaves were isolated from 2/4-months old plants which were grown in a greenhouse (at 25°C, under a 16 h photoperiod, using white fluorescent light with a maximum intensity of 70 µmol m^{-2} s^{-1}).

The explants were surface sterilized using 70% ethanol for 30 sec and 30% of commercial bleach for 20-30 min. After being rinsed with sterile water five times, they were used for bacterial inoculation.

Leaf segments with the lower epidermis upwards were placed on an agar-solidified, hormone-free half-strength B5 medium including vitamins (Gamborg), supplemented with 30 g L^{-1} sucrose, and were used for transformation with *A. tumefaciens* C58C1

strain carrying the pRiA4 of *A. rhizogenes*, the binary vector pGreenII35S-CrTF12 and pSoup. Infections with *A. rhizogenes* A4 wild type and *A. tumefaciens* C58C1 carrying the pRiA4 of *A. rhizogenes* wild-type served as negative control. In order to investigate the effect of the pGreenII35S vector by itself on the hairy roots cultures, infection with *A. tumefaciens* C58C1 strain carrying the pRiA4 of *A. rhizogenes*, the empty binary vector pGreenII35S and pSoup has also been performed. The infections were made by wounding the leaves on the midrib segment with a sterile hallow needle loaded with 48 hour-old bacterial colonies. After 48 h of co-cultivation, the leaves were maintained on the half-strength B5 solid medium, supplemented with 30 g L^{-1} sucrose containing 500 mg L^{-1} cefotaxime (Duchefa) to eliminate bacterial contamination. The hairy roots appeared 2–4 weeks after the infection and they were excised and cultured individually. The hairy root lines were grown at 25°C in the dark and were routinely subcultured to fresh half-strength B5 solid medium, supplemented with 30 g L^{-1} sucrose every two weeks. The cefotaxime concentration was successively reduced along 3 months.

2.6 Analysis of Hairy Roots

2.6.1 Genomic DNA extraction

Genomic DNA was extracted from the putative engineered hairy root lines and wild-type hairy root lines according to the CTAB extraction protocol, described in Appendix 5.

2.6.2 Polymerase chain reaction analysis for *vir* and *rolC* genes

The presence of the *vir* genes in root tissues was analysed by PCR, in order to monitor the elimination of the *Agrobacterium*. Once proven the complete elimination of bacteria, by the absence of the amplification of the *vir* genes, it is necessary to demonstrate the insertion of the *rolC* genes in the plant genome. The primers used for amplification of the *vir* gene and the *rolC* genes, as well as, some of their most relevant parameters are shown in Table 20.

Table 20. Primers used to detect *vir* and *rolC* genes.

Primer	Sequence (5' to 3')	Size (bp)	Tm (°C)
VIR I	ATG TCG CAA GGC AGT AAG CCC A	22	61.5
VIR II	GGA GTC TTT CAG CAT GGA GCA A	22	57.5
Rol_C1	TAA CAT GGC TGA AGA CGA CC	20	54.7
Rol_C2	AAA CTT GCA CTC GCC ATG CC	20	59.1

The composition of the PCR reactions is the typical (Table 5) and the thermal cycling conditions are shown in Table 21.

Table 21. PCR conditions performed for the detection of *vir* and *rol* genes.

***vir* genes**						
	95°C	95°C	58°C	72°C	72°C	Hold 6°C
Time	5 min	45 sec	45 sec	4 min	5 min	∞
Number of cycles	1	35			1	
***rolC* genes**						
	95°C	95°C	58°C	72°C	72°C	Hold 6°C
Time	5 min	1 min	30 sec	1 min	5 min	∞
Number of cycles	1	35			1	

2.7 Subcellular localization in *C. roseus* mesophyll protoplasts

2.7.1 Preparation of plasmid DNA for PEG-mediated transformation of protoplasts

Preparation of pDNA for protoplast transformation, i.e. ultrapure and highly concentrated, was performed using the Plasmid Midi-Prep Kit (Qiagen), according to the manufacturer's instructions but with some modifications. A starting culture of 100 mL was used, and the bacterial cells were harvested by centrifugation at 5000 rpm (3850 g) for 15 min at 4°C (Universal 320R, Hettich). The alkaline lysis procedure was carried out in 50 mL Falcon tubes and the cleaning of the lysate was done by centrifuging at 5000 rpm (3850 g) for 60 min at 4°C followed by another centrifugation under the same conditions, but for 30 min. The equilibration of the column, the DNA binding, washing and elution steps were all performed as indicated in the original protocol. Isopropanol precipitation was performed in 15 mL Falcon tubes, and the solution was then divided into 2 mL eppendorf tubes and centrifuged at maximum speed (13000 rpm / 15700 g) for 45 minutes at 4°C (Centrifuge 5415 R, Eppendorf). The resulting pellets were washed with 500 µL of 70% ethanol and centrifuged again under the same conditions, for 5 minutes. The pellets were allowed to dry at RT and then resuspended in the appropriated volume of sterile H_2O. The resulting pDNA integrity was checked by agarose gel electrophoresis as described above. pDNA quantification was performed spectrophotometrically using a NanoDrop ND1000 device. To sum up the pDNA preparations obtained through this method, gravity-flow with QIAGEN anion-exchange tips from the QIAGEN Plasmid Midi Prep kit, result in endotoxin-free, ultrapure and highly concentrated DNA, which is a key factor to

improving transformation efficiency. pDNA preps isolated in such a way typically render a yield of ~1.5-2 g L^{-1}.

2.7.2 Isolation of *C.roseus* mesophyll protoplasts

C. roseus mesophyll protoplasts were isolated according to Duarte et al. (2011). Leaves of the second or third pair of adult plants were cut into ~1 mm strips, after excising the central vein, and were immediately transferred to a Petri dish with 5 mL of digestion medium composed of 2 % (w/v) cellulose (Onozuka R-10, Duchefa), 0.3 % (w/v) macerozyme (Onozuka R-10, Serva) and 0.1 % pectinase (Sigma) dissolved in MM buffer (0.4 M mannitol and 20 mM Mes, pH 5.6-5.8), keeping the abaxial face down. The material was vacuum infiltrated for 15 min with 30 sec pulses and then incubated at 25°C, in the dark, during 3 h. After this incubation, the Petri dishes were placed on an orbital shaker (~60 rpm) for 15 min in the dark and at RT to help release the protoplasts. The suspension was filtered through a 100 µm nylon mesh and the filtrate was transferred into 15 mL falcon tubes. To pellet the protoplasts, the suspension was centrifuged at 65 g for 5 min at 20 °C. The supernatant was removed, the protoplasts were washed twice in MM buffer and once in cold W5 solution (154 mM NaCl, 125 mM CaCl 2.2H 2O, 5 mM KCl and 2 mM Mes, pH 5.7), and the pellet was resuspended in 2 mL of W5. Protoplasts were counted using a haemocytometer and were incubated on ice for 30 min. After that time, the suspension was once again centrifuged and the pellet was resuspended in an appropriate volume of MMg buffer (0.4 M mannitol, 15 mM MgCl2 and 4 mM Mes, pH 5.7) to yield a final protoplast concentration of $5x10^6$ cells.mL^{-1}.

2.7.3 PEG-mediated transformation of *C. roseus* mesophyll protoplasts

C .roseus mesophyll protoplasts were transformed following the procedure by Duarte et al. (2011). Typically, 20 µg of pDNA (Midi-prep) were mixed with 100 µL of protoplast suspension in a 2 mL round bottom eppendorf. One volume (110 µL) of PEG solution (40 % w/v PEG, 0.2 M mannitol and 0.1 M CaCl 2.2H 2O) was added drop by drop to this mixture, flicking the tube after every drop. The tubes were left to incubate for 15 min at RT and then four volumes of W5 solution (440 µL) were slowly added, flicking the tube after every three drops. The mixture was centrifuged at 56 g for 2 min, with acceleration and deceleration set at the minimum. The supernatant was removed, the pellet was gently resuspended in 100 µL of W5 solution, the protoplasts were

transferred to 15 mL falcon tubes containing 900 µL of W5, and were incubated in the dark at 25°C, with the tubes lying in a slight slope, for at least 2 days.

2.7.4 Confocal microscopy

After 24 h and 48 h of incubation, fluorescence inside the protoplasts was examined using a Leica SP2 AOBS SE confocal microscope equipped with a scanhead with an argon laser. Visualization of GFP was performed using an excitation wavelength of 488 nm and an emission wavelength window from 506 to 538 nm. Visualization of chloroplast auto fluorescence was performed using the same excitation wavelength and an emission wavelength window from 648 to 688 nm.

3 Results

3.1 Molecular cloning of the *C. roseus* novel transcription factor candidate genes

3.1.1 *In silico* retrieval of the full cDNA sequence of the *C. roseus* novel transcription factors

Considering that the major regulatory mechanism behind alkaloid production in plants is thought to be performed by transcription factors (TFs), it was hypothesized that TFs differentially expressed in *C. roseus* idioblasts could constitute good candidate genes for the activation of the expression of biosynthetic genes and/or transporters involved in the late, bottleneck TIA pathway. These novel *C. roseus* transcription factors (CrTFs) may ultimately enable genetic manipulation resulting in an increase of TIA levels, namely of the anticancer VLB and VCR.

Previously, differential transcriptomic profiling of idioblasts by cDNA-AFLP enabled the identification of several TF tags differentially expressed in this cell type. The TF tags were subjected to a BLAST search against the Medicinal Plant Genomics Resource database (MPGR, http://medicinalplantgenomics.msu.edu/index.shtml), which holds an extensive collection of transcriptomic data from *C. roseus*, sampled across a diverse set of developmental tissues, in cultured cells and in roots following elicitor treatment (Gongora-Castillo et al., 2012). This enabled to retrieve the full coding sequence for 6 novel CrTFs that were chosen to be isolated and cloned. Table 22 represents the accession number and BLASt results for the six selected CrTFs.

Table 22. Accession numbers of the six novel CrTFs selected for isolation and cloning. Results of BLAST search in the MPGR database of the initial cDNA-AFLP tags for each CrTF is also represented.

Accession	Hit Score	E value	Best HSP Cov	Best HSP Ident	Short name
cra_locus_12026_iso_4_len_5472_ver_3	1235	6.6e-50	96.18%	99.21%	CrTF12
cra_locus_517_iso_2_len_3386_ver_3	1446	8.3e-60	96.75%	98.67%	CrTF19
cra_locus_5366_iso_3_len_1889_ver_3	1516	7.7e-64	83.50%	93.60%	CrTF79
cra_locus_4787_iso_1_len_2989_ver_3	1967	2.1e-84	97.54%	99.50%	CrTF152
cra_locus_4594_iso_2_len_2637_ver_3	2778	5.5e-121	98.76%	99.46%	CrTF194
cra_locus_5262_iso_1_len_2814_ver_3	344	4.7e-13	69.49%	92.68%	CrTF246

Throughout this work, (March 2013) two more *C. roseus* transcriptomic databases, became available: PhytoMetaSyn (http://www.phytometasyn.ca/) (Xiao et al., 2013) and Cathacyc (http://www.cathacyc.org/) (Van Moerkercke et al., 2013).

In order to confirm that the sequences referred in Table 22 were complete and correct, BLAST searches were also performed using the PhytoMetaSyn database. The obtained sequences were run at the ORF Finder platform (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). The relevant information provided by the two different databases, as well as the alignments of the nucleotide sequences and of the translated sequences is shown in Appendix 6. One of de tags that was initially retrieved from MPGR as an incomplete coding sequence, CrTF61, became also available as a full length sequence after BLAST against PhytoMetaSyn. The alignment of the nucleotide sequences provided by the two databases and some additionally information can be found in Appendix 7. However, as the complete nucleotide sequence of CrTF61 only became available very late in the time frame of this work, its isolation and cloning were not attempted.

3.1.2 Primary cloning of the *C. roseus* novel transcription factors

The full length sequences for CrTF12, CrTF19, CrTF79, CrTF152, CrTF194 and CrTF246, all obtained from the BLAST against the MPGR database, were used to design primers to amplify the full coding sequences of the candidate genes through RT-PCT. CrTF152, CrTF194 and CrTF246 were successfully amplified using with RNA obtained from leaves as a starting material (Figure 8).

To amplify CrTF12, CrTF19 and CrTF79 it was necessary to use as starting material RNA isolated from leaf protoplasts, naked cells obtained after digestion of the cell wall - (Figure 6) - from 4 month-old plants (results not shown). The PCR amplification products of CrTF12, CrTF19, CrTF79, CrTF194 and CrTF246 were purified (Figure 9) and used for ligation with the vector pGEM®-T Easy, to try to obtain a stable source of DNA for the subsequent sub-cloning tasks.

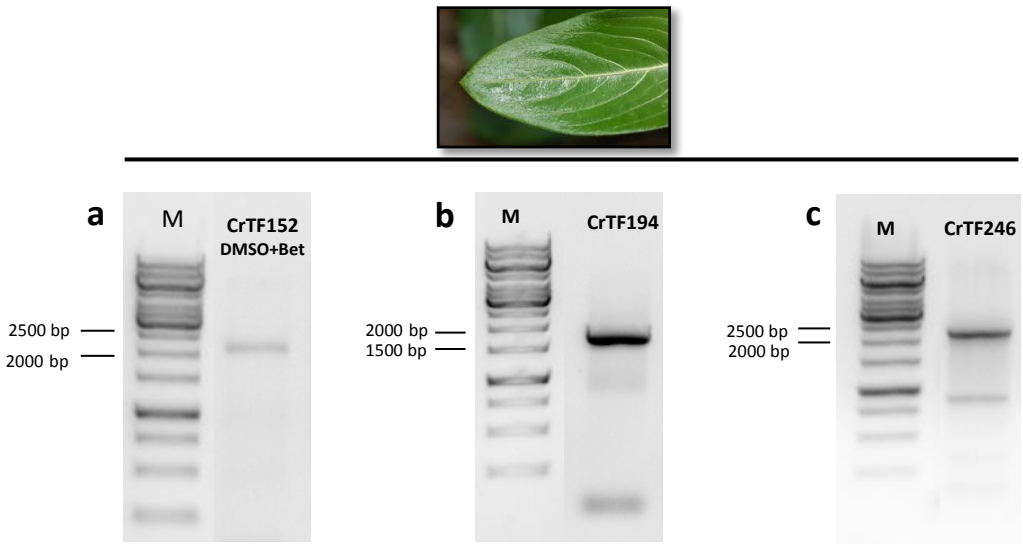


Figure 8. Amplification of CrTF152, CrTF194 and CrTF246 for subcloning into pGEM®-T Easy vector. a) CrTF152 amplicon – 2090 bp; b) CrTF194 amplicon – 1812 bp; c) CrTF246 amplicon – 2184 bp. These images correspond to non-contiguous lanes of the same gel; M, Molecular weight marker, Generuler ™ DNA Ladder Mix (Thermo Scientific); DMSO, dimethyl sulfoxide ; Bet, betaine; 1% agarose gel stained with ethidium bromide.

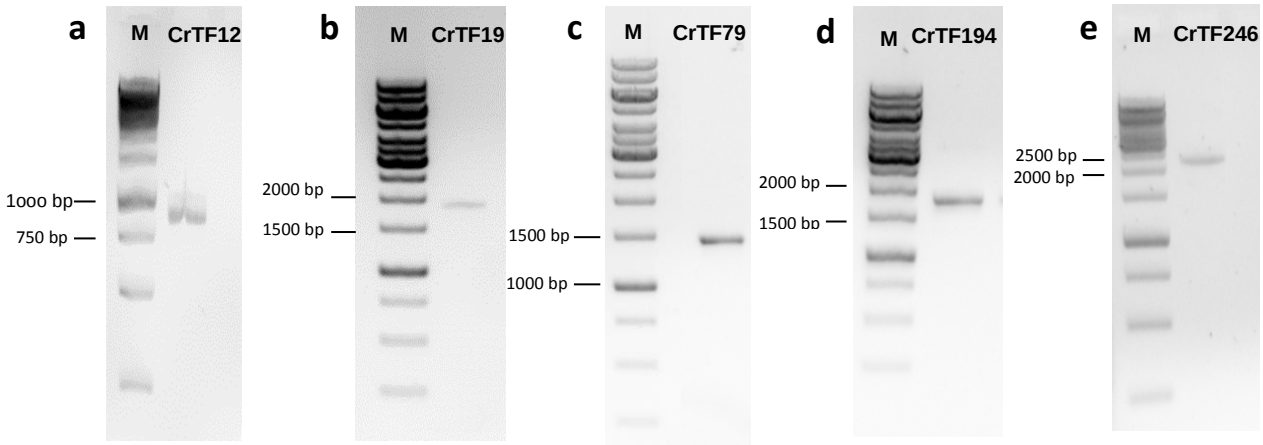


Figure 9. DNA recovered from the agarose gel after amplification of the selected CrTFs for subcloning into pGEM®-T Easy vector. a) CrTF12 amplicon – 867 bp; b) CrTF19 amplicon – 1920 bp; c) CrTF79 amplicon – 1452 bp; d) CrTF194 amplicon – 1812 bp; e) CrTF246 amplicon – 2184 bp. These images correspond to non-contiguous lanes of the same gel; M, Molecular weight marker, Generuler TM DNA Ladder Mix (Thermo Scientific); 1% agarose gels stained with ethidium bromide.

It was not possible to amplify CrTF152 in enough amounts to try the primary cloning in pGEM®-T Easy. A specific band with the correct size was only obtained once, in spite of all the attempts to optimize the amplification reaction/conditions shown in Table 4 of Material and Methods.

E. coli DH5α competent cells were transformed with the products of ligation of pGEM®-T Easy with the coding sequences for CrTF12, CrTF19, CrTF79, CrTF194 and CrTF246. The pDNA of a few colonies putatively transformed with each CrTF cDNA, except for CrTF246, was submitted to restriction analysis, in order to select the positive ones (Figure 10 and Figure 11).

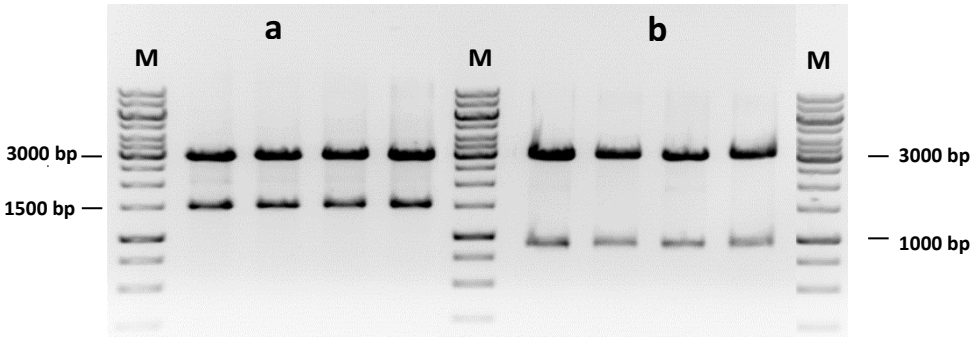


Figure 10. Restriction analysis of pGEM-T easy–CrTF79 and pGEM-T easy–CrTF12. a) All the represented clones are positive, which is translated by the presence of a band with ~3000 bp corresponding to the plasmid and a band with~1500 bp (CrTF79 amplicon – 1452 bp); b) All the represented clones are positive, which is translated by the presence of a band with ~3000 bp corresponding to the plasmid and a band with~1000 bp (CrTF12 amplicon - 867 bp); These images correspond to non-contiguous lanes of the same gel ; M, Generuler ™ DNA Ladder Mix (Thermo Scientific); 1% agarose gels stained with ethidium bromide.

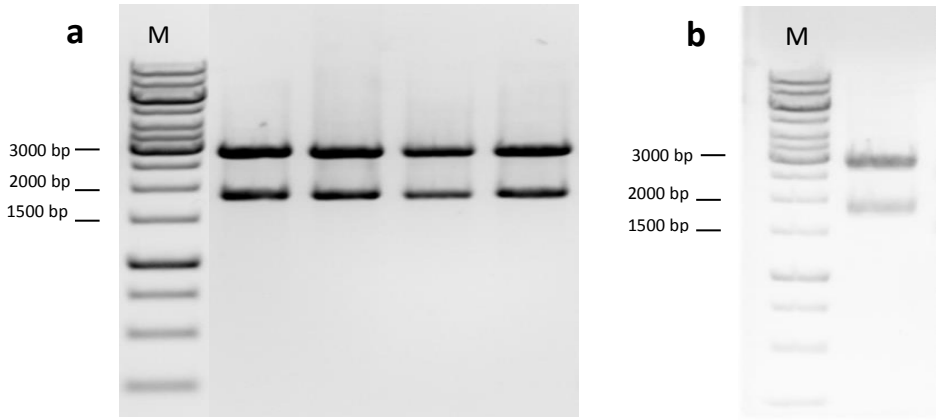


Figure 11. Restriction analysis of pGEM-T easy–CrTF19 and pGEM-T easy–CrTF194. a) All the represented clones are positive, which is translated by the presence of a band with ~3000 bp corresponding to the plasmid and a band with~2000 bp (CrTF19 amplicon – 1920 bp); b) The represented clone is positive, which is translated by the presence of a band with ~3000 bp corresponding to the plasmid and a band with~2000 bp (CrTF19 amplicon – 1812 bp); These images correspond to non-contiguous lanes of the same gel ; M, Generuler ™ DNA Ladder Mix (Thermo Scientific); 1% agarose gels stained with ethidium bromide.

Positive colonies were retrieved for every CrTF except for CrTF246. The whole cloning process was reinitiated for CrTF246, but, this time it was not possible to obtain a PCR amplification product using the conditions that had worked before. Touch-down PCR under different enhancing conditions was tried in order to amplify CrTF246 but no positive results were obtained, as is shown in Figure 12.

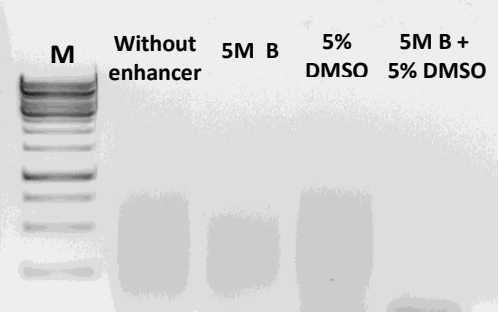


Figure 12. Amplification of CrTF246 using TD-PCR and several enhancers. M, Generuler ^{TM} DNA Ladder Mix (Thermo Scientific); Bet, betaine; DMSO, dimethyl sulfoxide; 1% agarose gels stained with ethidium bromide.

For each construct, at least two positive colonies identified by restriction analysis were sequenced. The resulting sequences were aligned with the transcripts retrieved from the MPGR and PhytoMetaSyn databases. For CrTF12, a perfect alignment was obtained, as shown in (Figure 13), making this clone fitted for further characterization. In the case of the CrTF79 sequence, a mismatch in one nucleotide was detected as shown in (Figure 14). However, alignment of the amino-acid sequences of their translated proteins showed that the detected mismatch did not produce any difference in the protein sequence, and should correspond to a single nucleotide polymorphism of the plant population used (Figure 15). This conclusion was also supported by the fact that the same mismatch was present in all the CrTF79 clones sequenced. Therefore, this clone was also selected for further work. CrTF19 and CrTF194 were both stranded at this point, because of the significant differences between the expected sequences, retrieved from the MPGR database, and the sequences obtained for the cloned cDNAs. The alignments obtained are shown in Appendix 6.

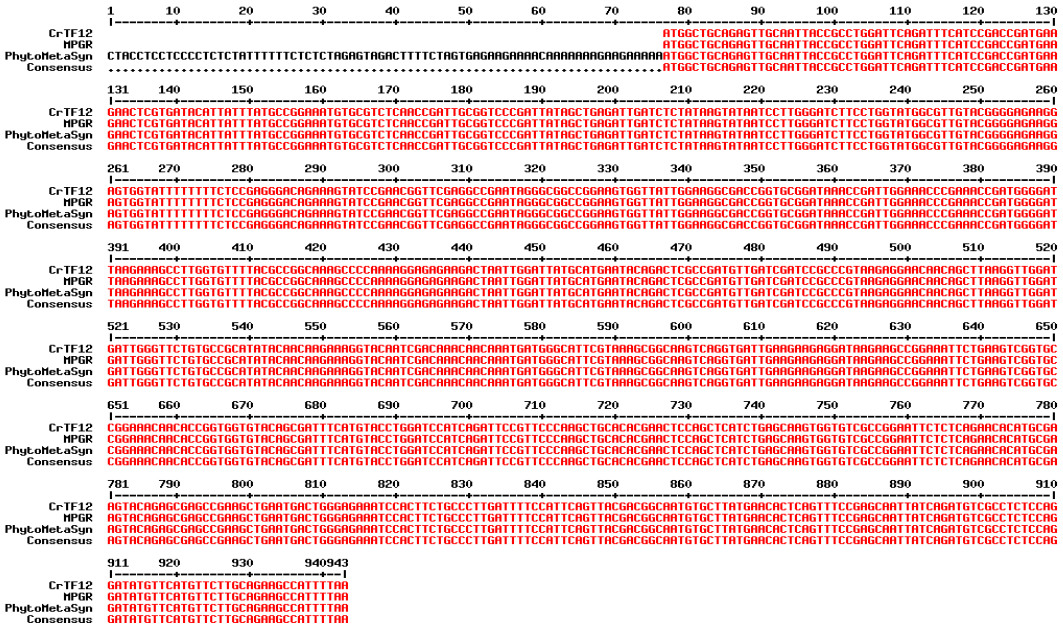


Figure 13. Alignment between the *CrTF12* nucleotide sequence retrieved from MPGR and PhytoMetaSyn databases and the sequence of the cloned CrTF12. Red corresponds to regions common to all aligned sequences.

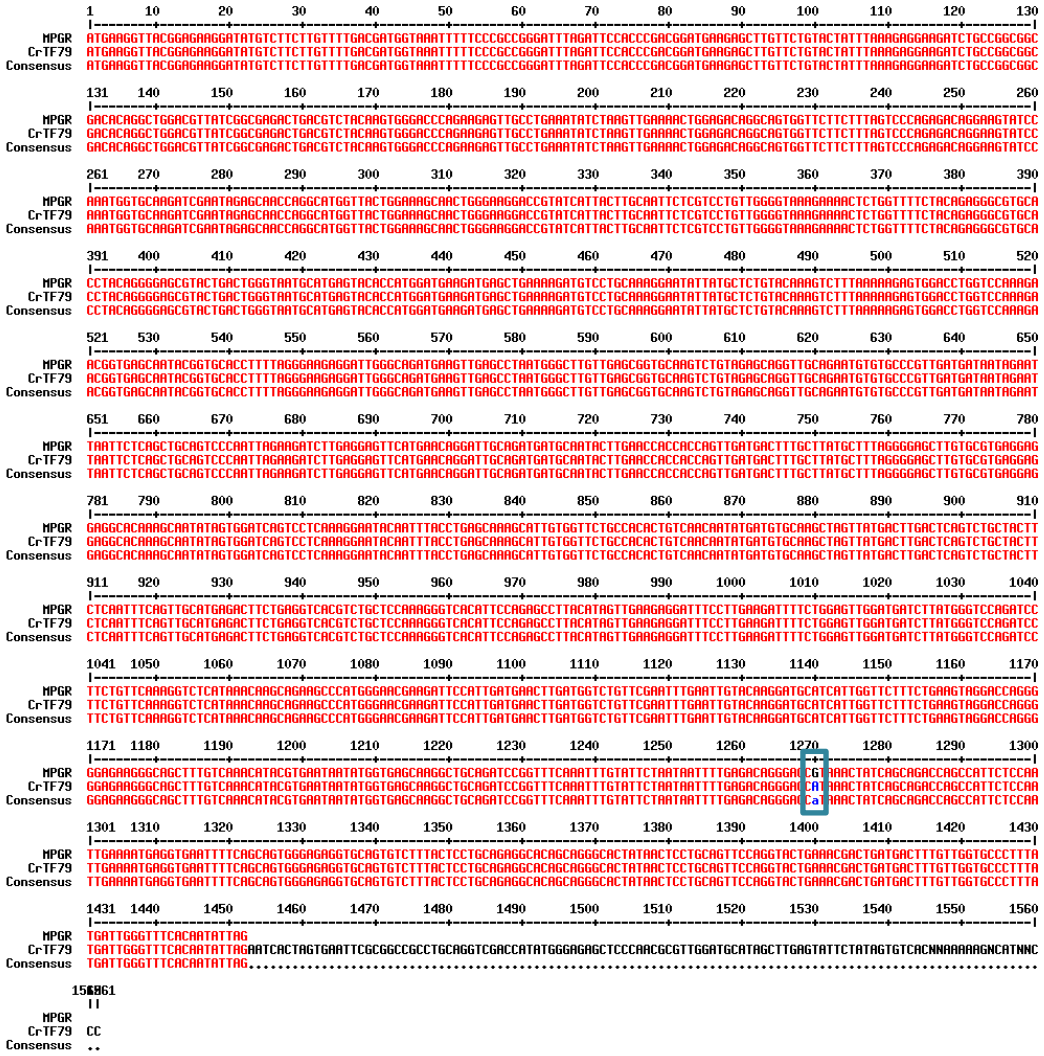


Figure 14. Alignment between the *CrTF79* nucleotide sequence retrieved from MPGR database and the sequence of the cloned CrTF79. Blue corresponds to regions with differences between the aligned sequences; red corresponds to regions common to all aligned sequences.

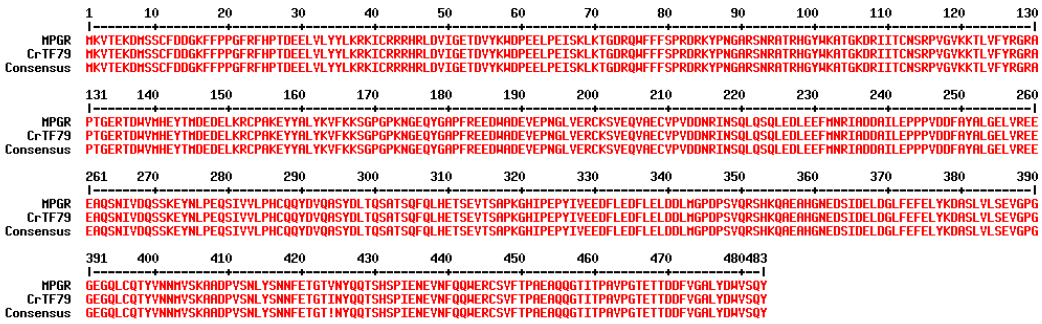


Figure 15. Alignment between the CrTF79 protein sequence retrieved from MPGR database and the translate DNA sequence of CrTF79 cloned in pGEM-T Easy; Red corresponds to regions common to all aligned sequences.

3.1.3 Sub-cloning of the *C. roseus* novel transcription factors for *Agrobacterium* mediated plant transformation

Work was then carried on with the CrTF12 and the CrTF79 cDNAs, which were sub-cloned into the binary vector pGreensII35S, in other to obtain *Agrobacterium*-mediated transgenic hairy roots overexpressing these CrTFs. For this, the pDNA harbouring the CrTF12 and CrTF79 clones was used as template to amplify the CrTFs full coding sequences using primers that added the appropriate restriction sites to allow directional sub-cloning in pGreenII35S. The sequences were successfully amplified using the primers described in Table 3 and the conditions described in Table 8 and Table 9 of Material and Methods (Figure 16). To perform the ligation between the CrTF12 and CrTF79 coding sequences flanked by the appropriated restriction sites and the binary vector pGreensII35S, both plasmid and inserts were digested with *Not*I and *Sgs*I (*Asc*I) restriction enzymes and the digested products were recovered from an agarose gel as described above. The purified products obtained are shown in Figure 17.

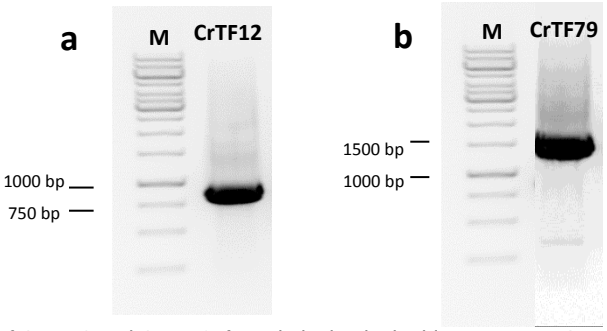


Figure 16. Amplification of CrTF12 and CrTF79 for subcloning in the binary vector pGreenII35S. a) CrTF12 amplicon – 867 bp; b) CrTF79 amplicon – 1452 bp. These images correspond to non-contiguous lanes of the same gel; M, Generuler ™ DNA Ladder Mix (Thermo Scientific); 1% agarose gels stained with ethidium bromide.

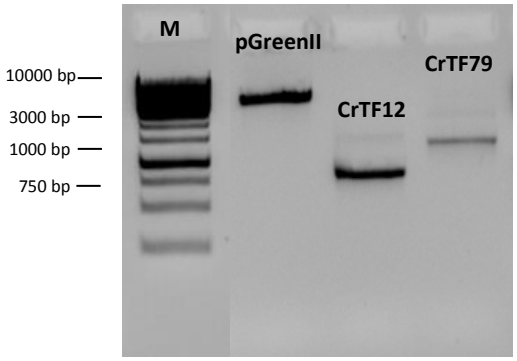


Figure 17. DNA recovered from agarose gel after restriction with *Sal*I and *Not*I restriction enzymes of the vector pGreensII35S and the amplified CrTF12 and 79 cDNAs. pGreenII35S - ~ 8000 bp; CrTF12 – 867 bp; CrTF79 – 1452 bp. These images correspond to non-contiguous lanes of the same gel; M**,** Generuler ™ DNA Ladder Mix (Thermo Scientific); 1% agarose gels stained with ethidium bromide.

Competent *E. coli* DH5α cells were transformed with the ligation reactions (Table 14) and the pDNA of at least four colonies for each CrTF was extracted and submitted to restriction digestion. As shown in Figure 18 all the selected colonies were positive, and at least two positive colonies for each CrTF were sequenced. The resulting *CrTF12* sequences were aligned with the sequences retrieved from the MPGR and PhytoMetaSyn databases and a perfect alignment has been obtained, as shown in Figure 19.

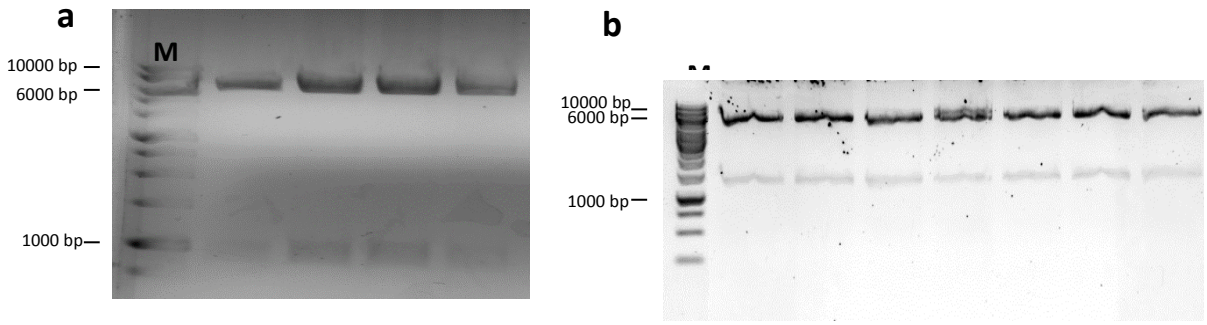


Figure 18. Restriction analysis of pGreenII35S – CrTF12 and pGreenII35S – CrTF79. a) All the represented clones are positive, which is translated by the presence of a band with ~8000 bp corresponding to the plasmid and a band with ~1000 bp corresponding to the CrTF12 amplicon (867 bp); b) All the represented clones are positive, which is translated by the presence of a band with ~8000 bp corresponding to the plasmid and a band with ~ 1500 bp corresponding to the CrTF79 amplicon (1452 bp). M, Generuler ^{TM} DNA Ladder Mix (Thermo Scientific); 1% agarose gels stained with ethidium bromide**.**

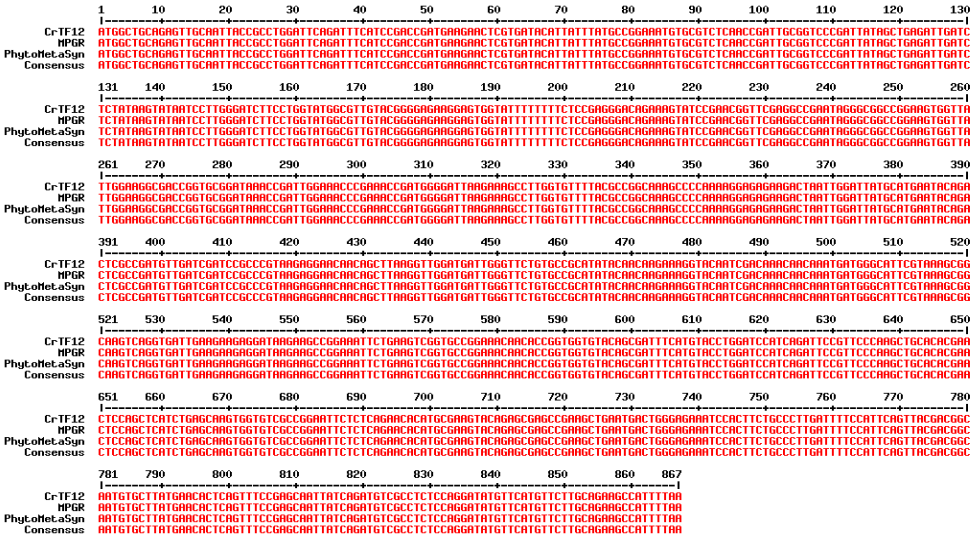


Figure 19. Alignment between the *CrTF12* nucleotide sequence retrieved from MPGR and PhytoMetaSyn databases and the sequence of CrTF12 cloned in pGreenII35S. Red corresponds to regions common to all aligned sequences.

However, all the CrTF79 positive clones showed sequences differing from each other, and with significant differences from the sequence obtained during primary cloning and the MPGR sequence. The sub-cloning process was repeated, but there were again many changes in the sub-cloned sequences. This work will proceed using a different binary vector.

3.1.4 Sub-cloning of the *C. roseus* novel transcription factors for subcellular localization

The *CrTF12* sequence was further sub-cloned in the expression vector pTH2 to generate a *CrTF12-GFP* (green fluorescent protein) fusion gene enabling to investigate the subcellular localization of this TF. The fusion was generated with the *CrTF12* gene fused at the N-terminus coding sequence of GFP, in order to enable the putative nuclear N-terminus signal of the expressed CrTF12 to be freely available for recognition.

Once more, the pGEM-Teasy-CrTF12 construct was used as template to amplify the TF full coding sequence using primers that added the appropriate restriction sites to allow directional cloning in the pTH2 vector. The amplification products and the DNA recovered from the agarose gel are shown in Figure 20.

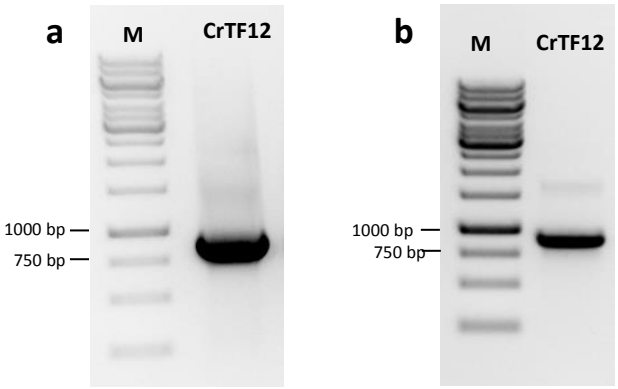


Figure 20. a) Amplification of CrTF12 for subcloning in the expression vector pTH2; CrTF12 amplicon – 867 bp; b) CrTF12 recovered from the agarose gel; CrTF12 amplicon – 867 bp. M, Generuler ™ DNA Ladder Mix (Thermo Scientific); 1% agarose gels stained with ethidium bromide.

To perform the ligation between the CrTF12 coding sequence flanked by the appropriated restriction sites, and the expression vector pTH2, both plasmid and inserts were digested with *Sal*I and *Nco*I restriction enzymes, as described in (Table 15).

Both, pTH2 and the CrTF12 gene digested products were recovered from an agarose gel and the purified products obtained are shown in Figure 21.

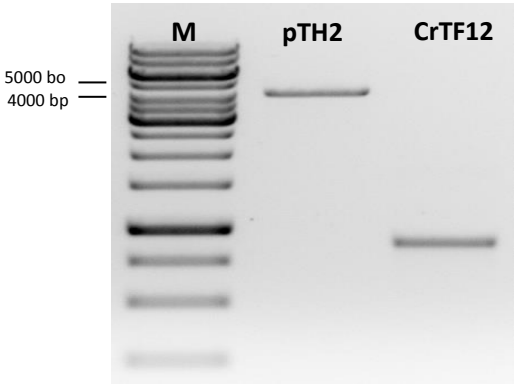


Figure 21. DNA recovered from agarose gel after restriction with *Sal*I and *Nco*I restriction enzymes (Thermo Scientific); pTH2 - ~ 3000 bp; CrTF12 – 867 bp; M**,** Generuler ™ DNA Ladder Mix (Thermo Scientific); 1% agarose gels stained with ethidium bromide.

E. coli TOP10 competent cells were transformed, the pDNA was extracted and the insertion of CrTF12 was confirmed by comparing the linearized empty pTH2 and the linearized pTH2 that presumably harboured the CrTF12. The insertion was confirmed by the difference of approximately 1Kb, corresponding to the molecular weight of CrTF12 (Figure 22). All tested colonies were positive for the presence of CrTF12.

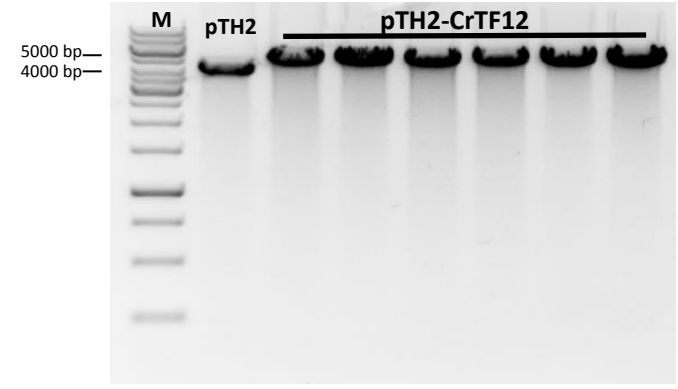


Figure 22. Restriction analysis of the expression vector pTH2 and the pTH2 – CrTF12 to linearize the plasmid. The insertion was confirmed by the difference of approximately 1Kb, corresponding to the molecular weight of CrTF12 – 867 bp. M, Generuler ™ DNA Ladder Mix (Thermo Scientific); 1% agarose gels stained with ethidium bromide.

3.2 Generation of *C. roseus* hairy roots overexpressing CrTF12

The binary vector pGreensII35S harbouring the coding sequence of CrTF12 together with the helper plasmid pSoup were used to transform *Agrobacterium tumefaciens* C58C1 strain carrying the pRiA4 of *A. rhizogenes* (from here onwards referred as *A. tumefaciens* C58C1). *A. tumefaciens* C58C1 transformed with the empty pGreensII35S and the pSoup helper was used as a negative control. The obtained colonies were used to infect *C.roseus* leaves, leading to the generation of hairy roots from the leaf explants. As additional controls, infection of *C. roseus* leaves was also performed with the non transformed *A. tumefaciens* C58C1 and with *A. rhizogenes* A4.

Hairy roots infected with both wild type strains *A. tumefaciens* C58C1 and *A. rhizogenes* A4, as well as *A. tumefaciens* C58C1 harbouring the empty pGreenII35S or pGreensII35S-CrTF12, have been successfully obtained. Lines resulting from the infections with the wild type strains showed the typical growth profile expected for hairy roots: extensive proliferation from the infected plant wound, modified gravitropic response, reduced apical dominance and increased lateral branching (Figure 23).

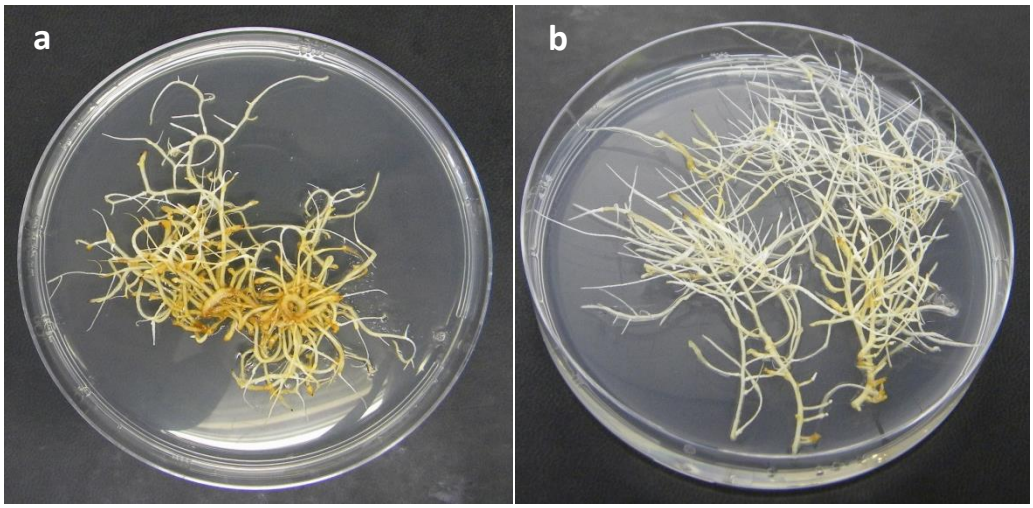


Figure 23. Aspect of the hairy root lines that resulted from the infections with the wild-type strains of: a) *A. tumefaciens* C58C1 and b) *A. rhizogenes* A4.

On the other hand, lines infected with the bacteria harbouring pGreenII35S empty or carrying CrTF12 genes, showed a very slow growth (Figure 24). In an attempt to increase the growth rate of these lines, two strategies were followed. Some samples were grown in liquid half strength B5 medium with gentle shaking (70 rpm) and other samples were grown in solid half strength B5 medium, supplemented with 2 $mg.L^{-1}$ of IBA (the later induces root generation). Both strategies proved to be ineffective, since the growth rate was not increased in any of the cases.

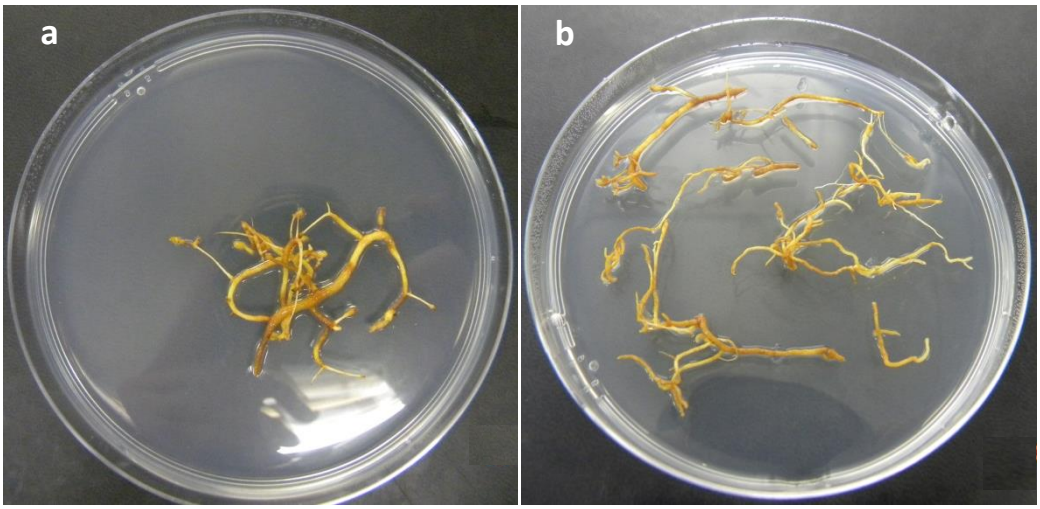


Figure 24. Hairy root lines transformed with *A. tumefaciens* C58C1 harbouring a) the empty pGreenII35S and pSoup and b) pGreenII35S-CrTF12 and pSoup. These lines show a retarded growth rate.

3.3 Analysis of *C. roseus* hairy root lines

In order to assess if the obtained *C. roseus* hairy root lines are indeed transgenic, the presence of the *vir* and *rolC* genes was tested. Ideally, hairy root lines are ready for subsequent analysis when the *vir* genes are absent, indicating *Agrobacterium* elimination, and when the *rolC* genes are present, indicating the transference of the T-DNA genes to the plant.

The presence of one of the *vir* genes in root tissues was assayed by PCR over genomic DNA (gDNA), in order to prove the elimination of *Agrobacterium.* Only if this has happened, will the amplification of the transgenic gene be able to confirm plant transformation. Until now, the total elimination of the bacteria was only proven for one line transformed with the wild-type *A. rhizogenes* A4 strain (Line A) (Figure 25). As a positive control for the PCR reaction, the DNA of *A. tumefaciens* C58C1 was used. As expected, in the positive control a band with 438 bp was amplified corresponding to the presence of the target *vir* gene (Figure 25).Together with another wild type line, D, transformed with *A. rhizogenes* A4 strain, two putative CrTF12 overexpression transgenic lines, B and C, that reached enough biomass to be tested were still positive for the presence of the *vir* gene (Figure 25).Therefore, its transgenic status could not be confirmed yet.

Once proven the complete elimination of the bacteria by the absence of the amplification of the *vir* genes, we carried the analysis of line A onto the demonstration of the insertion of the *rolC* gene in the plant genome. In the wild-type hairy-root line A the *rolC* specific PCR yielded the expected product (534 bp). Again, as a positive

control for the PCR reaction, the DNA of the *A. tumefaciens* C58C1 was used (Figure 26).

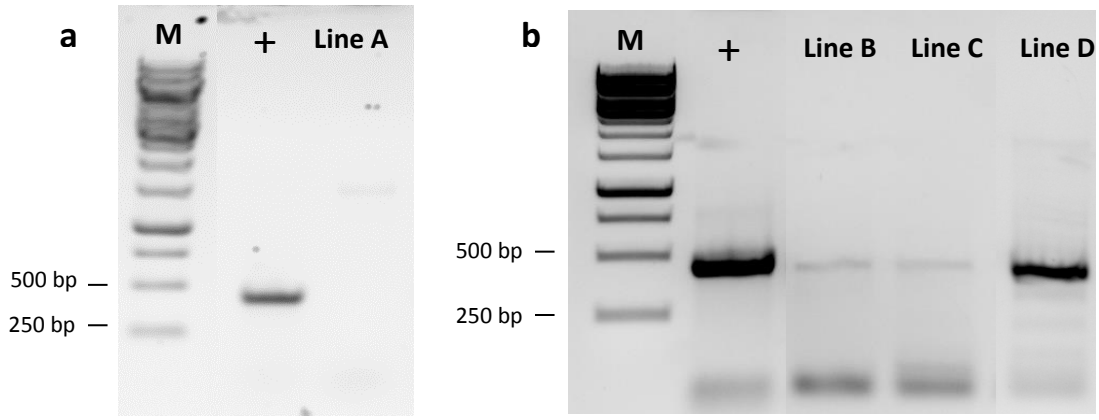


Figure 25. a) PCR analysis for the presence of a *vir* gene (438 bp) in the wild-type hairy root line A; b) PCR analysis for the presence of a *vir* gene (438 bp) in the putative CrTF12 overexpression transgenic hairy roots lines B and C and in the wild-type hairy root line ; +, positive control using DNA from *A. tumefaciens* C58C1. These images correspond to non-contiguous lanes of the same gel M, Generuler ™ DNA Ladder Mix (Thermo Scientific); 1% agarose gel stained with ethidium bromide.

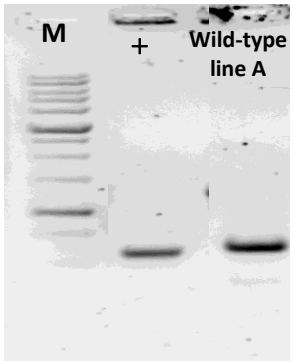


Figure 26. PCR analysis for the presence of *rolC* gene (534 bp) in the wild-type hairy root line A; +, positive control using DNA from *A. tumefaciens* C58C1. These images correspond to non-contiguous lanes of the same gel M, Generuler ™ DNA Ladder Mix (Thermo Scientific); 1% agarose gel stained with ethidium bromide.

These results indicated that the *C. roseus* hairy root line A contained the *rolC* genes in its genome, indicating successful transformation, and that it was free of bacteria as shown by the inability to detect the *vir* gene. However, this was not possible to confirm yet for the lines putatively transformed with the empty pGreenII35S vector or with pGreenII35S-CrTF12, since the two tested still presented bacterial growth (Figure 25) and most of the lines generated could not even be tested yet, due to their slow growth rate (Figure 24).

In this work it was expected to obtain enough biomass of transgenic *C. roseus* hairy root lines to assess the TIA biosynthetic gene expression levels by qPCR, and to

analyze, through HPLC, the TIA levels / profile present in this expression system, to detect the impact of CrTF12 overexpression. This work is currently underway.

3.4 Subcellular localization of the *C. roseus* novel transcription factors

3.4.1 *In silico* prediction of subcellular localization

An *in silico* prediction of the subcellular localization of all six selected CrTFs (Table 22) was performed using the TargetP 1.1 tool (http://www.cbs.dtu.dk/services/TargetP/) and WoLFPSORT (http://www.genscript.com/psort/wolf_psort.html).

Table 23 *In silico* subcellular localization analysis for the CrTFs from *C. roseus* under study, using the TargetP1.1 tool and the WoLFPSORT. cTP, chloroplast transit peptide; mTP, mitochondria targeting peptide; SP, signal peptide; Cyto, cytosol; Nuc, nucleus; Cysk, cytoskeleton; Chlor, chloroplast; Mito, mitochondria; Pero, peroxisomes; Plas, plasma membrane. .

	TargetP 1.1				WoLF PSORT						
	cTP	**mTP**	**SP**	**Other**	**Cyto**	**Nuc.**	**Cysk**	**Chlor.**	**Mito**	**Pero**	**Plas.**
CrTF12	0.292	0.072	0.043	0.548	4	8	2	-	-	-	-
CrTF19	0.229	0.193	0.034	0.617	-	8	-	5	-	-	-
CrTF79	0.316	0.118	0.020	0.504	-	14	-	-	-	-	-
CrTF152	0.218	0.137	0.036	0.479	-	6	-	5	2	-	-
CrTF194	0.166	0.209	0.034	0.582	-	10	-	-	-	3	-
CrTF246	0.299	0.120	0.078	0.417	3	8	-	-	-	-	1

Analysis with TargetP 1.1 indicated that there is a probability for all the TFs to be localized in the nucleus, since the values associated with 'other compartments' group, where nucleus is included, are the highest. When using the WoLF Psort tool, the results of *in silico* analysis clearly point to a nuclear localization of all TFs.

3.4.2 Subcellular localization of a CrTF12-GFP fusion in *C. roseus* mesophyll protoplasts

To investigate the subcellular localization of CrTF12, the TF gene for which work had progressed the most, *CrTF12*, was sub-cloned in the expression vector pTH2, to create an N-terminal fusion with GFP as depicted in Figure 27.

Figure 27. Schematic representation of the construct used to investigate the subcellular localization of CrTF12. 35S, cauliflower mosaic virus strong promoter; nos T, *Agrobacterium tumefaciens* nopaline synthase terminator..

Then pTH2-CrTF12 plasmid DNA was used to transiently transform *C. roseus* mesophyll protoplasts and GFP fluorescence was observed under the confocal microscope 24h and 48h after the transformation. No differences were observed in the

fluorescence pattern between the 24h and 48 h time-points. In both, GFP fluorescence could clearly be observed in the nucleus, confirming a nuclear localization for CrTF12, as expected and predicted *in silico* (Figure 28). The presence of a non-transformed cell in the Figure 28 lower panel, with a complete absence of green fluorescence in the nucleus, proves that the observed nuclear green fluorescence is specific to the GFP fusion.

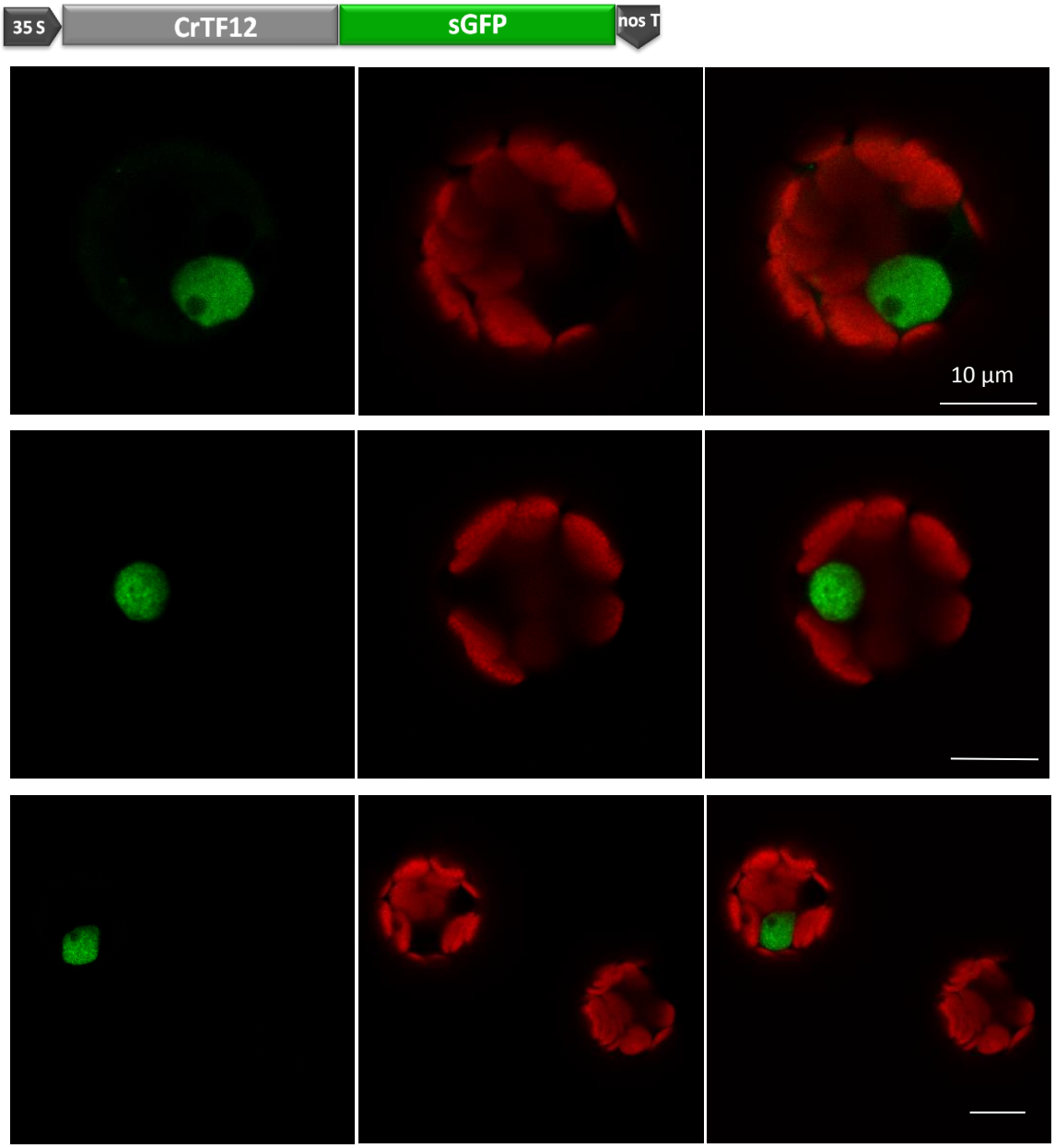


Figure 28. Transient expression of 35S::sGFP-CrTF12 in *C. roseus* mesophyll protoplasts observed under the confocal microscope 48h after the transformation. Left - GFP fluorescence; Middle – Chloroplasts autofluorescence; Right – merged images.

4 Discussion

In this work, we performed the *in silico* analysis of six transcription factor genes previously shown to be differentially expressed in TIA accumulating idioblast cells from *C. roseus*: *CrTF12, CrTF19 CrTF79, CrTF152, CrTF194* and *CrTF246*. We successfully isolated five of them (excluding *CrTF152*), and were able to clone four (excluding *CrTF246*). *CrTF12*, a NAC TF gene overexpressed in idioblasts was successfully used for *Agrobacterium*-mediated transformation of *C. roseus* leaves to generate putatively transgenic hairy roots, which are expected to overexpress *CrTF12*. *In silico* analysis predicted for all the CrTFs a nuclear localization, which was proved to be true for CrTF12 using GFP-tagging and transient expression in *C. roseus* mesophyll protoplasts. A discussion of the most important aspects and results of this work is provided below.

4.1 Transcriptomic databases as a valuable tool

Catharanthus, as the single biological source of the anticancer TIAs, VCR and VLB, has become one of the most studied medicinal plants (El-Sayed and Verpoorte, 2007). However, since these compounds accumulate in very low amounts in the plant and are difficult to extract, metabolic engineering approaches have been used to increase specialized metabolite levels in plants, or alternative *in vitro* production systems. One major barrier to the implementation of scientifically directed approaches aiming to manipulate the yield and spectrum of pharmacologically active compounds in plants is the limited information on the regulatory networks, enzymes and transporters involved in the production and flux of these compounds (Gongora-Castillo et al., 2012). Complementing this limitation, there is a rising number of transcriptomic databases with information on medicinal plants allowing the visualization and interpretation of large-scale omic-data (Gongora-Castillo et al., 2012).

In the case of *C. roseus*, whose genome has not been sequenced yet, these databases are particularly valuable, since the use of transcriptomes and/or proteomes can circumvent the necessity for fully assembled and annotated genomes. The databases MPGR and PhytoMetaSyn were used in this work to do the BLAST search of the transcription factor sequence tags differentially expressed in idioblasts, in order to try to retrieve their full length coding sequences. The postulated hypothesis is that one of these CrTFs could eventually constitute a master switch in the regulation of the TIA pathway. Although the information provided by databases is invaluable, it may present

errors and therefore it should not be assumed as certain. There are several most probable explanations for these observations. First of all, there are some common problems frequently associated with the databases information, including sequence artefacts, poor quality reads and primer/adapter contamination (Xiao et al., 2013). Furthermore, to construct a database such as those mentioned above, a huge volume of information has to be treated, namely in means of contigs generation and sequencing, thus making it likely for the occurrence of some errors. Finally, we need to remember that since the transcripts shown in the databases are the longest representative isoform of a cluster, it is possible that similar sequences are lost in the process, namely the existence of allele polymorphisms (Gongora-Castillo et al., 2012; Xiao et al., 2013).

For example the CrTF19 sequence presents a (questionable) duplication of 41 nucleotides in the retrieved sequence which is not present in the sequences obtained in this work. Most probably, this also constitutes an error in the contig of the MPGR database, since this duplication is unlikely to occur, and in fact, later we were able to prove that it was absent in the sequence obtained from PhytoMetaSyn. The one base difference observed for the isolated CrTF79 when compared to the corresponding sequence retrieved from the MPGR database may also constitute a database error, since the same difference appeared in all of the sequenced clones. However, this difference did not alter the translated protein sequence, suggesting that it may rather be a single nucleotide polymorphism between the different *C. roseus* varieties/populations used here and by MPGR.

Overall, the transcriptomic databases represent a valuable tool for the unravelling of metabolic pathways that have not been characterized yet, however these data have to be analysed with caution for the above cited reasons.

4.2 Amplification and cloning of the *C. roseus* novel transcription factor candidate genes

The polymerase chain reaction (PCR) is a powerful tool to investigate the mechanisms underlying life. However, its optimization requires the adjustment of a variety of parameters. The PCR conditions, the use of enhancers as well as the integrity of the RNA / cDNA are extremely important aspects to consider that could be the cause of difficulties in the amplification, especially in plants enriched in secondary metabolites.

In this work, different sources of RNA have been used, in an attempt to amplify all the CrTFs under study. Since the main aim of this work was to study CrTFs up-regulated in

idioblasts, ideally the best way to isolate them would be to use RNA extracted from idioblasts as initial template. However, since the isolation of these cells is expensive, laborious and demands the use of equipment that does not exist at IBMC, it is rarely performed, and only for very specific aims. Therefore, the best option was to try the amplification using RNA from leaves and / or protoplasts obtained from plants at different developmental stages. During the course of this work, RNA with better yield and quality was consistently obtained whenever using protoplasts, most probably because of the absence of the cell wall. Thus, from then onwards, it was decided to always isolate protoplasts instead of using whole leaves to extract RNA. As an alternative, RNA extracted from seedlings (one month old plants) was also used, with the reasoning that at the first stages of development there is a bigger chance of a major number of genes being expressed, since cell differentiation is still incomplete. This strategy proved successful only once for the amplification of CrTF246.

The intense PCR amplification trials performed in this work enabled to conclude that the integrity of *C. roseus* RNA and cDNA is extremely vulnerable and, most of the times, the extraction of RNA, reverse transcription (RT) and PCR amplification have to be performed in an immediate sequence in order to achieve success. On the other hand, the use of PCR enhancers proved to be less efficient than expected. Only in the case of CrTF152 an increase in the intensity of the amplification band was observed using 5% DMSO and 1 M betaine. However, this result was not reproducible in further attempts. Since the different trials were made at different times of the year, a seasonal influence in gene expression must not be excluded. For all the primers, several annealing temperatures were tested, in order to obtain the highest amplification efficiency. Overall, the integrity of the cDNA, the rapidness of the whole process and the chosen annealing temperature were proven to be the most relevant aspects to achieve a successful DNA amplification in the course of this work.

Regarding the cloning / transformation processes, one of the approaches that could make a difference is the cleaning of the ligation reactions (through a column), a procedure that is not usually done in our lab, to get rid of the ligase, etc, since this was shown to make transformation more efficient. Moreover, most of the transformations were performed with *E. coli* DH5α while the use of *E. coli* TOP10 competent cells could possibly enable to generate colonies for the unsuccessful cloning attempts. Contrary to DH5α, the galE gene is mutated in the TOP10 strain, resulting in shorter chains of lipopolysaccharides (LPS) on the cell surface. These shorter chains are thought to be the reason for the greater transformation efficiencies that can be reached with TOP10

cells, and other cells with this genotype, since long LPS chains may block the entrance sites of the pDNA (van Die et al., 1984).

4.3 Generation of transgenic hairy roots

In this work, transgenic *C. roseus* hairy roots putatively overexpressing CrTF12 have been obtained from Agrobacterium-transformed leaf explants. Moreover, infection with the wild-type strains *A. tumefaciens* C58C1 carrying the pRiA4 and *A. rhizogenes* A4 were performed, rendering transgenic hairy roots carrying the *rol* genes.

A huge difference has been observed in the growth rate between the hairy roots obtained with the *Agrobacterium* wild type strains, and with *A. tumefaciens* C58C1 carrying either empty pGreenII35S vector, or pGreenII35S-CrTF12, both together with pSoup. The reduced growth rate of the hairy roots is most probably due to the presence of the pGreenII35S itself rather than to overexpression of CrTF12, since both hairy roots transformed with the empty version of the vector and with the vector carrying the CrTF12 coding sequence showed a similar abnormally slow growth rate, when compared to the hairy roots obtained using the wild type *Agrobacterium* strains. In order to circumvent this reduced growth rate problem, the first thing to be done would be to do the cloning of the CrTFs in the binary vector pCAMBIA2301, which has been previously used with success in *C. roseus* (Suttipanta et al., 2011; Wang et al., 2012).

The growth rate of hairy roots infected with *A. rhizogenes* or *A. tumefaciens* wild-type strains was very similar and no significantly differences were detected suggesting that either of the strains used are adequate to this methodology.

4.4 Subcelular localization of CrTFs

By definition, transcription factors are regulatory proteins that modulate the expression of specific groups of genes typically by sequence-specific DNA binding. They can act as activators or repressors of gene expression, mediating either an increase or a decrease in the accumulation of messenger RNA (Broun, 2004). Given the above, TFs are usually expected to be present in the nucleus. However, there are reports of TFs that are localized in the cytosol and are only transported to the nucleus (or vice-versa) after activation by a stress signal (Mukai et al., 1996; Taylor et al., 1996; Pandey et al., 2004). In any case, determining the subcellular localization of putative TFs is important to confirm and understand their potential as transcriptional regulators.

A first approach to the analysis of the subcellular localization of the *CrTF* genes was performed *in silico*. According to TargetP1.1 and WoLFPSORT, all of the CrTFs under study have the highest probability to be localized in the nucleus (Table 23). Since the greatest progress in this work was achieved with CrTF12, this TF emerged as the obvious candidate for experimental characterization of its subcellular localization. With this purpose, a CrTF12-sGFP fusion construct was generated. Only the N-terminal fusion of CrTF12 with sGFP was generated, since nuclear localization signals (NLS) are localized at the N-terminal region of a nuclear protein. A C-terminal version of this construct would muffle the NLS. The transient expression of the *CrTF12-GFP* fusion construct in *C. roseus* mesophyll protoplasts showed a clear nuclear localization of the green fluorescence from the fusion protein. Our results suggest that this is the subcellular localization of CrTF12 *in planta*, however it has to be considered that the GFP fluorescence in the nucleus may be due to the translocation of an otherwise cytosolic protein to the nucleus, dependent on the protoplasting and transformation method processes, known to be highly stressful for the cells.

5 Conclusions

This work has initiated the isolation, cloning and characterization of a set of transcription factors (CrTFs) differentially expressed in the leaf idioblast cells from *C. roseus*, which are specialized in the late, bottleneck biosynthetic steps of the anticancer alkaloids VLB and VCR. The identification of these *CrTF* candidate genes constitutes a major breakthrough in the study of the transcriptional regulatory network modulating the complex metabolism of TIAs in *C. roseus* leaves, and may include regulators of the late steps of TIA biosynthesis and of the transmembrane TIA transporters. This study set the conditions for the successful amplification of 5 novel *CrTF* genes, performed the molecular cloning of four of those *CrTFs* that are now ready for further characterization, and most likely established a transgenic model for the functional characterization of one of the novel CrTFs, which subcellular localization was also determined. Overall, this study as launched the bases for future work aiming at disclosing the full transcriptional network regulating the TIA pathway, whose knowledge may ultimately lead to the successful manipulation of *C. roseus* for higher yields of the anticancer alkaloids.

6 Bibliography

A. Hoekema PRH (1983) A binary plant vector strategy based on separation of vir- and T-region of the Agrobacterium tumefaciens Ti-plasmid. Nature 303: 179 - 118

Alting-Mees MA, Short JM (1989) pBluescript II: gene mapping vectors. Nucleic Acids Res 17: 9494

Bai Y, Pattanaik S, Patra B, Werkman JR, Xie CH, Yuan L (2011) Flavonoid-related basic helix-loop-helix regulators, NtAn1a and NtAn1b, of tobacco have originated from two ancestors and are functionally active. Planta 234: 363-375

Batra J, Dutta A, Singh D, Kumar S, Sen J (2004) Growth and terpenoid indole alkaloid production in Catharanthus roseus hairy root clones in relation to left- and right-termini-linked Ri T-DNA gene integration. Plant Cell Rep 23: 148-154

Bennouna J, Campone M, Delord JP, Pinel MC (2005) Vinflunine: a novel antitubulin agent in solid malignancies. Expert Opin Investig Drugs 14: 1259-1267

Bhadra R (1994) Establishment, cultivation and optimization of Hairy Roots of *Catharanthus roseus* for the synthesis of Indole Alkaloids. Houston, Texas

Bourgaud F, Gravot A, Milesi S, Gontier E (2001) Production of plant secondary metabolites: a historical perspective. Plant Science 161: 839-851

Brown S, Renaudin JP, Prevot C, Guern J (1984) Flow-cytometry and sorting of plant-protoplasts - technical problems and physiological results from a study of ph and alkaloids in *Catharanthus roseus* Physiologie Vegetale 22: 541-554

Butelli E, Titta L, Giorgio M, Mock H-P, Matros A, Peterek S, Schijlen EGWM, Hall RD, Bovy AG, Luo J, Martin C (2008) Enrichment of tomato fruit with health-promoting anthocyanins by expression of select transcription factors. Nature Biotechnology 26: 1301-1308

Carqueijeiro I (2013) Unravelling the metabolism and transmembrane transport of the highly valuable medicinal alkaloids from Catharanthus roseus PhD Thesis. Faculty of Sciences, University of Porto

Carqueijeiro I, Noronha H, Duarte P, Geros H, Sottomayor M (2013) Vacuolar Transport of the Medicinal Alkaloids from Catharanthus roseus Is Mediated by a Proton-Driven Antiport. Plant Physiol 162: 1486-1496

Chang AC, Cohen SN (1978) Construction and characterization of amplifiable multicopy DNA cloning vehicles derived from the P15A cryptic miniplasmid. J Bacteriol 134: 1141-1156

Chatel G, Montiel G, Pre M, Memelink J, Thiersault M, Saint-Pierre B, Doireau P, Gantet P (2003) CrMYC1, a Catharanthus roseus elicitor- and jasmonate-responsive bHLH transcription factor that binds the G-box element of the strictosidine synthase gene promoter. Journal of Experimental Botany 54: 2587-2588

Citovsky V, Lee L-Y, Vyas S, Glick E, Chen M-H, Vainstein A, Gafni Y, Gelvin SB, Tzfira T (2006) Subcellular localization of interacting proteins by bimolecular fluorescence complementation in planta. Journal of Molecular Biology 362: 1120-1131

Costa MMR, Hilliou F, Duarte P, Pereira LG, Almeida I, Leech M, Memelink J, Barcelo AR, Sottomayor M (2008) Molecular cloning and characterization of a vacuolar class III peroxidase involved in the metabolism of anticancer alkaloids in *Catharanthus roseus*. Plant Physiology 146: 403-417

De Carolis E, Chan F, Balsevich J, De Luca V (1990) Isolation and Characterization of a 2-Oxoglutarate Dependent Dioxygenase Involved in the Second-to-Last Step in Vindoline Biosynthesis. Plant Physiol 94: 1323-1329

De Luca V, Cutler AJ (1987) Subcellular Localization of Enzymes Involved in Indole Alkaloid Biosynthesis in Catharanthus roseus. Plant Physiol 85: 1099-1102

De Luca V, St Pierre B (2000) The cell and developmental biology of alkaloid biosynthesis. Trends Plant Sci 5: 168-173

Duarte P, Ribeiro D, Henriques G, Hilliou F, Rocha AS, Lima F, Amorim I, Sottomayor M (2011) Cloning and characterization of a candidate gene from the medicinal plant *Catharanthus roseus* through transient expression in mesophyll protoplasts. *In* GG Brown, ed, Molecular Cloning – Selected Applications in Medicine and Biology. In Tech

El-Sayed M, Verpoorte R (2007) *Catharanthus* terpenoid indole alkaloids: Biosynthesis and regulation. Phytochemistry Reviews 6: 277-305

Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300: 1005-1016

Facchini P, Bird D, St-Pierre B (2004) Can *Arabidopsis* make complex alkaloids? Trends in Plant Science 9: 116-122

Facchini PJ (2001) Alkaloid biosynthesis in plants: Biochemistry, cell biology, molecular regulation, and metabolic engineering applications. Annual Review of Plant Physiology and Plant Molecular Biology 52: 29-66

Facchini PJ, Luca VD (2008) Opium poppy and Madagascar periwinkle: model non-model systems to investigate alkaloid biosynthesis in plants. Plant Journal 54: 763-784

Feng JX, Liu D, Pan Y, Gong W, Ma LG, Luo JC, Deng XW, Zhu YX (2005) An annotation update via cDNA sequence analysis and comprehensive profiling of developmental, hormonal or environmental responsiveness of the Arabidopsis AP2/EREBP transcription factor gene family. Plant Mol Biol 59: 853-868

Framond AJd (1983) Mini–Ti: A New Vector Strategy for Plant Genetic Engineering. Nature Biotechnology 1: 262 - 269

Gelvin S (2003) *Agrobacterium*-Mediated Plant Transformation: the Biology behind the "Gene-Jockeying" Tool. Microbiology and Molecular Biology Reviews 67: 16-37

Gelvin SB (2009) Agrobacterium in the genomics age. Plant Physiol 150: 1665-1676

Giddings (2011) Discovery, characterization, and rational design of the enzymes involved in monoterpene indole alkaloid biosynthesis in Madagascar periwinkle. PhD. Massachusetts Institute of Technology

Gigant B, Wang C, Ravelli RB, Roussi F, Steinmetz MO, Curmi PA, Sobel A, Knossow M (2005) Structural basis for the regulation of tubulin by vinblastine. Nature 435: 519-522

Giri A, Narasu ML (2000) Transgenic hairy roots. recent trends and applications. Biotechnol Adv 18: 1-22

Goddijn OJ, de Kam RJ, Zanetti A, Schilperoort RA, Hoge JH (1992) Auxin rapidly down-regulates transcription of the tryptophan decarboxylase gene from Catharanthus roseus. Plant Mol Biol 18: 1113-1120

Goddijn OJ, Pennings EJ, van der Helm P, Schilperoort RA, Verpoorte R, Hoge JH (1995) Overexpression of a tryptophan decarboxylase cDNA in Catharanthus roseus crown gall calluses results in increased tryptamine levels but not in increased terpenoid indole alkaloid production. Transgenic Res 4: 315-323

Gomez A, Lopez JA, Pintos B, Camafeita E, Bueno MA (2009) Proteomic analysis from haploid and diploid embryos of Quercus suber L. identifies qualitative and quantitative differential expression patterns. Proteomics 9: 4355-4367

Gong W, Shen YP, Ma LG, Pan Y, Du YL, Wang DH, Yang JY, Hu LD, Liu XF, Dong CX, Ma L, Chen YH, Yang XY, Gao Y, Zhu D, Tan X, Mu JY, Zhang DB, Liu YL, Dinesh-Kumar SP, Li Y, Wang XP, Gu HY, Qu LJ, Bai SN, Lu YT, Li JY, Zhao JD, Zuo J, Huang H, Deng XW, Zhu

YX (2004) Genome-wide ORFeome cloning and analysis of Arabidopsis transcription factor genes. Plant Physiol 135: 773-782

Gongora-Castillo E, Childs KL, Fedewa G, Hamilton JP, Liscombe DK, Magallanes-Lundback M, Mandadi KK, Nims E, Runguphan W, Vaillancourt B, Varbanova-Herde M, Dellapenna D, McKnight TD, O'Connor S, Buell CR (2012) Development of transcriptomic resources for interrogating the biosynthesis of monoterpene indole alkaloids in medicinal plant species. PLoS One 7: e52506

Guirimand G, Burlat V, Oudin A, Lanoue A, St-Pierre B, Courdavault V (2009) Optimization of the transient transformation of *Catharanthus roseus* cells by particle bombardment and its application to the subcellular localization of hydroxymethylbutenyl 4-diphosphate synthase and geraniol 10-hydroxylase. Plant Cell Reports 28: 1215-1234

Guirimand G, Courdavault V, Lanoue A, Mahroug S, Guihur A, Blanc N, Giglioli-Guivarc'h N, St-Pierre B, Burlat V (2010) Strictosidine activation in Apocynaceae: Towards a "nuclear time bomb"? BMC Plant Biology: 182

Guirimand G, Guihur A, Ginis O, Poutrain P, Héricourt F, Oudin A, Lanoue A, St-Pierre B, Burlat V, Courdavault V (2011) The subcellular organization of strictosidine biosynthesis in *Catharanthus roseus* epidermis highlights several trans-tonoplast translocations of intermediate metabolites. FEBS Journal 278: 749-763

Guirimand G, Guihur A, Poutrain P, Héricourt F, Mahroug S, St-Pierre B, Burlat V, Courdavault V (2011) Spatial organization of the vindoline biosynthetic pathway in *Catharanthus roseus*. Journal of Plant Physiology 168: 549-557

Gutterson N, Reuber TL (2004) Regulation of disease resistance pathways by AP2/ERF transcription factors. Curr Opin Plant Biol 7: 465-471

Hanahan D, Jessee J, Bloom FR (1991) Plasmid transformation of Escherichia coli and other bacteria. Methods Enzymol 204: 63-113

Hartmann T (2007) From waste products to ecochemicals: fifty years research of plant secondary metabolism. Phytochemistry 68: 2831-2846

Hellens R, Mullineaux P, Klee H (2000) A guide to *Agrobacterium* binary Ti vectors. Trends in Plant Science 5: 446-451

Hilliou F, Christou P, Leech MJ (1999) Development of an efficient transformation system for *Catharanthus roseus* cell cultures using particle bombardment. Plant Science 140: 179-188

Hirata K MK, Miura Y (1994) Catharanthus roseus L. (Periwinkle): Production of Vindoline and Catharanthine in Multiple Shoot Cultures, Vol 26. Springer Berlin Heidelberg

Hong SB, Peebles CA, Shanks JV, San KY, Gibson SI (2006) Expression of the Arabidopsis feedback-insensitive anthranilate synthase holoenzyme and tryptophan decarboxylase genes in Catharanthus roseus hairy roots. J Biotechnol 122: 28-38

Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K (2007) WoLF PSORT: protein localization predictor. Nucleic Acids Res 35: W585-587

Junaid Aslam SHK, Zahid Hameed Siddiqui, Zohra Fatima, Mehpara Maqsood, Mukthar Ahmad Bhat, Sekh Abdul Nasim, Abdul Ilah, Iffat Zareen Ahmad, Saeed Ahmad Khan, Abdul Mujib, Sharma MP (2010) Catharanthus roseus (L.) G. Don. an important Drug: it's applications and production INTERNATIONAL JOURNAL OF COMPREHENSIVE PHARMACY 1: 1:16

Kato N, Dubouzet E, Kokabu Y, Yoshida S, Taniguchi Y, Dubouzet JG, Yazaki K, Sato F (2007) Identification of a WRKY protein as a transcriptional regulator of benzylisoquinoline alkaloid biosynthesis in Coptis japonica. Plant Cell Physiol 48: 8-18

Kingston DG (2009) Tubulin-interactive natural products as anticancer agents. J Nat Prod 72: 507-515

Levac D, Murata J, Kim WS, De Luca V (2008) Application of carborundum abrasion for investigating the leaf epidermis: molecular cloning of Catharanthus roseus 16-hydroxytabersonine-16-O-methyltransferase. Plant J 53: 225-236

Liu D-H, Ren W-W, Cui L-J, Zhang L-D, Sun X-F, Tang K-X (2011) Enhanced accumulation of catharanthine and vindoline in Catharanthus roseus hairy roots by overexpression of transcriptional factor ORCA2. African Journal of Biotechnology 10: 3260-3268

Liu DH, Jin HB, Chen YH, Cui LJ, Ren WW, Gong YF, Tang KX (2007) Terpenoid indole alkaloids biosynthesis and metabolic engineering in *Catharanthus roseus*. Journal of Integrative Plant Biology 49: 961-974

Liu L, White MJ, MacRae TH (1999) Transcription factors and their genes in higher plants functional domains, evolution and regulation. Eur J Biochem 262: 247-257

Loyola-Vargas VM, Galaz-Ávalos RM, Kú-Cauich R (2007) *Catharanthus* biosynthetic enzymes: The road ahead. Phytochemistry Reviews 6: 307-339

Mahroug S, Burlat V, St-Pierre B (2007) Cellular and sub-cellular organisation of the monoterpenoid indole alkaloid pathway in *Catharanthus roseus*. Phytochemistry Reviews 6: 363-381

Mano M (2006) Vinorelbine in the management of breast cancer: New perspectives, revived role in the era of targeted therapy. Cancer Treat Rev 32: 106-118

Memelink J, Gantet P (2007) Transcription factors involved in terpenoid indole alkaloid biosynthesis in *Catharanthus roseus*. Phytochemistry Reviews 6: 353-362

Memelink J, Verpoorte R, Kijne JW (2001) ORCAnization of jasmonate - responsive gene expression in alkaloid metabolism. Trends in Plant Science 6: 212-219

Menke FL, Champion A, Kijne JW, Memelink J (1999) A novel jasmonate- and elicitor-responsive element in the periwinkle secondary metabolite biosynthetic gene Str interacts with a jasmonate- and elicitor-inducible AP2-domain transcription factor, ORCA2. Embo J 18: 4455-4463

Mersey BG, Cutler AJ (1986) Differential distribution of specific indole alkaloids in leaves of *Catharanthus roseus*. Canadian Journal of Botany-Revue Canadienne De Botanique 64: 1039-1045

Morita M, Shitan N, Sawada K, Van Montagu MCE, Inze D, Rischer H, Goossens A, Oksman-Caldentey KM, Moriyama Y, Yazaki K (2009) Vacuolar transport of nicotine is mediated by a multidrug and toxic compound extrusion (MATE) transporter in Nicotiana tabacum. Proceedings of the National Academy of Sciences of the United States of America 106: 2447-2452

Mukai H, Miyahara M, Sunakawa H, Shibata H, Toshimori M, Kitagawa M, Shimakawa M, Takanaga H, Ono Y (1996) Translocation of PKN from the cytosol to the nucleus induced by stresses. Proc Natl Acad Sci U S A 93: 10195-10199

Murata J, Roepke J, Gordon H, De Luca V (2008) The leaf epidermome of *Catharanthus roseus* reveals its biochemical specialization. The Plant Cell 20: 524-542

Niwa Y, Hirano T, Yoshimoto K, Shimizu M, Kobayashi H (1999) Non-invasive quantitative detection and applications of non-toxic, S65T-type green fluorescent protein in living plants. The Plant Journal 18: 455-463

Noble RL (1990) The discovery of the vinca alkaloids--chemotherapeutic agents against cancer. Biochem Cell Biol 68: 1344-1351

O'Connor SE, Maresh JJ (2006) Chemistry and biology of monoterpene indole alkaloid biosynthesis. Nat Prod Rep 23: 532-547

Olsen AN, Ernst HA, Lo Leggio L, Skriver K (2005) NAC transcription factors: structurally distinct, functionally diverse. Trends in Plant Science 10: 79-87

Ouwerkerk PB, Memelink J (1999) Elicitor-responsive promoter regions in the tryptophan decarboxylase gene from Catharanthus roseus. Plant Mol Biol 39: 129-136

Palazon J, Cusido RM, Gonzalo J, Bonfill M, Morales C, Pinol MT (1998) Relation between the amount of rolC gene product and indole alkaloid accumulation in Catharanthus roseus transformed root cultures. Journal of Plant Physiology 153: 712-718

Pandey V, Mihara S, Fensome-Green A, Bolsover S, Cockcroft S (2004) Monomeric IgE stimulates NFAT translocation into the nucleus, a rise in cytosol Ca2+, degranulation, and membrane ruffling in the cultured rat basophilic leukemia-2H3 mast cell line. J Immunol 172: 4048-4058

Pasquali G, Goddijn OJ, de Waal A, Verpoorte R, Schilperoort RA, Hoge JH, Memelink J (1992) Coordinated regulation of two indole alkaloid biosynthetic genes from *Catharanthus roseus* by auxin and elicitors. Plant Mol Biol 18: 1121-1131

Pauw B, Hilliou FA, Martin VS, Chatel G, de Wolf CJ, Champion A, Pre M, van Duijn B, Kijne JW, van der Fits L, Memelink J (2004) Zinc finger proteins act as transcriptional repressors of alkaloid biosynthesis genes in Catharanthus roseus. J Biol Chem 279: 52940-52948

Peebles CA, Sander GW, Hughes EH, Peacock R, Shanks JV, San KY (2011) The expression of 1-deoxy-D-xylulose synthase and geraniol-10-hydroxylase or anthranilate synthase increases terpenoid indole alkaloid accumulation in Catharanthus roseus hairy roots. Metab Eng 13: 234-240

Pierre Potier NL, Yves Langlois and Françoise Guéritte (1975) Partial Synthesis of Vinblastine-Type Alkaloids. J.C.S. Chem. Comm: 670-671

Pollier J, Moses T, Goossens A (2011) Combinatorial biosynthesis in plants: a (p)review on its potential and future exploitation. Nat Prod Rep 28: 1897-1916

Qu LJ, Zhu YX (2006) Transcription factor families in Arabidopsis: major progress and outstanding issues for future research. Curr Opin Plant Biol 9: 544-549

Rao AQ, Bakhsh A, Kiani S, Shahzad K, Shahid AA, Husnain T, Riazuddin S (2009) The myth of plant transformation. Biotechnol Adv 27: 753-763

Roger P. Hellens EAE, Nicola R. Leyland, Samantha Bean and Philip M. Mullineaux (2000) pGreen: a versatile and flexible binary Ti vector for Agrobacterium-mediated plant transformation. Plant Molecular Biology 42: 819–832

Rushton PJ, Somssich IE, Ringler P, Shen QJ (2010) WRKY transcription factors. Trends Plant Sci 15: 247-258

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU (2005) A gene expression map of Arabidopsis thaliana development. Nat Genet 37: 501-506

Schroder G, Unterbusch E, Kaltenbach M, Schmidt J, Strack D, De Luca V, Schroder J (1999) Light-induced cytochrome P450-dependent enzyme in indole alkaloid biosynthesis: tabersonine 16-hydroxylase. FEBS Lett 458: 97-102

Shanks JV, Morgan J (1999) Plant 'hairy root' culture. Curr Opin Biotechnol 10: 151-155

Shoji T, Inai K, Yazaki Y, Sato Y, Takase H, Shitan N, Yazaki K, Goto Y, Toyooka K, Matsuoka K, Hashimoto T (2009) Multidrug and toxic compound extrusion-type transporters implicated in vacuolar sequestration of nicotine in tobacco roots. Plant Physiol 149: 708-718

Singh SN, Vats P, Suri S, Shyam R, Kumria MML, Ranganathan S, Sridharan K (2001) Effect of an antidiabetic extract of *Catharanthus roseus* on enzymic activities in streptozotocin induced diabetic rats. Journal of Ethnopharmacology 76: 269-277

Sottomayor M, dePinto M, Salema R, DiCosmo F, Pedreno M, AR B (1996) The vacuolar localization of a basic peroxidase isoenzyme responsible for the synthesis of alpha-3',4'-anhydrovinblastine in *Catharanthus roseus* (L) G. Don leaves. Plant Cell and Environment 19: 761-767

Sottomayor M, Lopes Cardoso I, Pereira L, Ros Barceló A (2004) Peroxidase and the biosynthesis of terpenoid indole alkaloids in the medicinal plant *Catharanthus roseus* (L.) G. Don. Phytochemistry Reviews 3: 159-171

Sottomayor M, Lopez-Serrano M, DiCosmo F, Ros Barcelo A (1998) Purification and characterization of alpha-3',4'-anhydrovinblastine synthase (peroxidase-like) from *Catharanthus roseus* (L.) G. Don. FEBS Letters 428: 299-303

Sottomayor M, Ros Barcelo A (2003) Peroxidase from *Catharanthus roseus* (L.) G. Don and the biosynthesis of alpha-3',4'-anhydrovinblastine: a specific role for a multifunctional enzyme. Protoplasma 222: 97-105

Sottomayor M, Ros Barceló A (2006) The Vinca Alkaloids: From biosynthesis and accumulation in plant cells, to uptake, activity and metabolism in animal cells. *In* Atta-ur-Rahman, ed, Studies in Natural Products Chemistry (Bioactive Natural Products), Vol 33. Elsevier Science Publishers, The Netherlands, pp 813-857

St-Pierre B, De Luca V (1995) A Cytochrome P-450 Monooxygenase Catalyzes the First Step in the Conversion of Tabersonine to Vindoline in Catharanthus roseus. Plant Physiol 109: 131-139

St-Pierre B, Laflamme P, Alarco AM, De Luca V (1998) The terminal O-acetyltransferase involved in vindoline biosynthesis defines a new class of proteins responsible for coenzyme A-dependent acyl transfer. Plant J 14: 703-713

St-Pierre B, Vazquez-Flota FA, De Luca V (1999) Multicellular compartmentation of *Catharanthus roseus* alkaloid biosynthesis predicts intercellular translocation of a pathway intermediate. The Plant Cell 11: 887-900

Suttipanta N (2011) CHARACTERIZATION OF G10H PROMOTER AND ISOLATION OF WRKY TRANSCRIPTION FACTORS INVOLVED IN CATHARANTHUS TERPENOID INDOLE ALKALOID BIOSYNTHESIS PATHWAY. Lexington, Kentucky

Suttipanta N, Pattanaik S, Kulshrestha M, Patra B, Singh SK, Yuan L (2011) The Transcription Factor CrWRKY1 Positively Regulates the Terpenoid Indole Alkaloid Biosynthesis in Catharanthus roseus. Plant Physiology 157: 2081-2093

Taylor GA, Thompson MJ, Lai WS, Blackshear PJ (1996) Mitogens stimulate the rapid nuclear to cytosolic translocation of tristetraprolin, a potential zinc-finger transcription factor. Mol Endocrinol 10: 140-146

Tzfira T, Kozlovsky SV, Citovsky V (2007) Advanced expression vector systems: New weapons for plant research and biotechnology. Plant Physiology 145: 1087-1089

van der Fits L, Hilliou F, Memelink J (2001) T-DNA activation tagging as a tool to isolate regulators of a metabolic pathway from a genetically non-tractable plant species. Transgenic Research 10: 513-521

van der Fits L, Memelink J (2000) ORCA3, a jasmonate-responsive transcriptional regulator of plant primary and secondary metabolism. Science 289: 295-297

van der Fits L, Memelink J (2000) ORCA3, a jasmonate-responsive transcriptional regulator of plant primary and secondary metabolism. Science 289: 295-297

van der Fits L, Memelink J (2001) The jasmonate-inducible AP2/ERF-domain transcription factor ORCA3 activates gene expression via interaction with a jasmonate-responsive promoter element. Plant J 25: 43-53

van der Fits L, Zhang H, Menke FL, Deneka M, Memelink J (2000) A Catharanthus roseus BPF-1 homologue interacts with an elicitor-responsive region of the secondary metabolite biosynthetic gene Str and is induced by elicitor via a JA-independent signal transduction pathway. Plant Mol Biol 44: 675-685

van der Heijden R, Jacobs DI, Snoeijer W, Hallared D, Verpoorte R (2004) The *Catharanthus* alkaloids: Pharmacognosy and biotechnology. Current Medicinal Chemistry 11: 607-628

van Die IM, Zuidweg EM, Bergmans JE, Hoekstra WP (1984) Transformability of galE variants derived from uropathogenic Escherichia coli strains. J Bacteriol 158: 760-761

Van Moerkercke A, Fabris M, Pollier J, Baart GJ, Rombauts S, Hasnain G, Rischer H, Memelink J, Oksman-Caldentey KM, Goossens A (2013) CathaCyc, a metabolic pathway database built from Catharanthus roseus RNA-Seq data. Plant Cell Physiol 54: 673-685

Vazquez-Flota F, De Carolis E, Alarco AM, De Luca V (1997) Molecular cloning and characterization of desacetoxyvindoline-4-hydroxylase, a 2-oxoglutarate dependent-dioxygenase involved in the biosynthesis of vindoline in Catharanthus roseus (L.) G. Don. Plant Mol Biol 34: 935-948

Verma P, Mathur AK, Srivastava A, Mathur A (2012) Emerging trends in research on spatial and temporal organization of terpenoid indole alkaloid pathway in Catharanthus roseus: a literature update. Protoplasma 249: 255-268

Vilas P. Sinkar FFW, Milton P. Gordon (1987) Molecular biology of Ri-plasmid—A review J. Biosci 11

47–57

Vom Endt D, Kijne JW, Memelink J (2002) Transcription factors controlling plant secondary metabolism: what regulates the regulators? Phytochemistry 61: 107-114

Wang Q, Xing S, Pan Q, Yuan F, Zhao J, Tian Y, Chen Y, Wang G, Tang K (2012) Development of efficient Catharanthus roseus regeneration and transformation system using agrobacterium tumefaciens and hypocotyls as explants. BMC Biotechnol 12: 34

Westekemper P WU, Gueritte F, Langlois N, Langlois N, Langlois Y, Potier P, Zenk MH (1980) Radioimmunoassay for the determination of the indole alkaloid vindoline in Catharanthus. Planta Medica 39: 24-23

Wink M (1993) The plant vacuole: a multifunctional compartment. Journal of Experimental Botany 44: 231-246

Wydro M, Kozubek E, Lehmann P (2006) Optimization of transient Agrobacterium-mediated gene expression system in leaves of *Nicotiana benthamiana*. Acta Biochimica Polonica 53: 289-298

Xiao M, Zhang Y, Chen X, Lee EJ, Barber CJ, Chakrabarty R, Desgagne-Penix I, Haslam TM, Kim YB, Liu E, MacNevin G, Masada-Atsumi S, Reed DW, Stout JM, Zerbe P, Zhang Y, Bohlmann J, Covello PS, De Luca V, Page JE, Ro DK, Martin VJ, Facchini PJ, Sensen CW (2013) Transcriptome analysis based on next-generation sequencing of non-model plants producing specialized metabolites of biotechnological interest. J Biotechnol 166: 122-134

Xiong Y, Liu T, Tian C, Sun S, Li J, Chen M (2005) Transcription factors in rice: a genome-wide comparative analysis between monocots and eudicots. Plant Mol Biol 59: 191-203

Yamasaki K, Kigawa T, Inoue M, Tateno M, Yamasaki T, Yabuki T, Aoki M, Seki E, Matsuda T, Tomo Y, Hayami N, Terada T, Shirouzu M, Tanaka A, Seki M, Shinozaki K, Yokoyama S (2005) Solution structure of an Arabidopsis WRKY DNA binding domain. Plant Cell 17: 944-956

Yanhui C, Xiaoyuan Y, Kun H, Meihua L, Jigang L, Zhaofeng G, Zhiqiang L, Yunfei Z, Xiaoxiao W, Xiaoming Q, Yunping S, Li Z, Xiaohui D, Jingchu L, Xing-Wang D, Zhangliang C, Hongya G, Li-Jia Q (2006) The MYB transcription factor superfamily of Arabidopsis: expression analysis and phylogenetic comparison with the rice MYB family. Plant Mol Biol 60: 107-124

Yoder LRaM, P. G. (1976) Reactions of Alkaloid and Histochemical Indicators in Laticifers and Specialized Parenchyma Cells of Catharanthus roseus (Apocynaceae). American Journal of Botany 63: 1167-1173

Zhang H, Hedhili S, Montiel G, Zhang Y, Chatel G, Pre M, Gantet P, Memelink J (2011) The basic helix-loop-helix transcription factor CrMYC2 controls the jasmonate-responsive expression of the ORCA genes that regulate alkaloid biosynthesis in Catharanthus roseus. The Plant Journal 67: 61-71

Zhang JZ (2003) Overexpression analysis of plant transcription factors. Curr Opin Plant Biol 6: 430-440

Zhao J, Dixon RA (2009) MATE transporters facilitate vacuolar uptake of epicatechin 3'-O-glucoside for proanthocyanidin biosynthesis in Medicago truncatula and Arabidopsis. Plant Cell 21: 2323-2340

Zhao J, Verpoorte R (2007) Manipulating indole alkaloid production by *Catharanthus roseus* cell cultures in bioreactors: From biochemical processing to metabolic engineering. Phytochemistry Reviews 6: 435-457

Zhou M-L, Hou H-L, Zhu X-M, Shao J-R, Wu Y-M, Tang Y-X (2011) Molecular regulation of terpenoid indole alkaloids pathway in the medicinal plant, Catharanthus roseus. Journal of Medicinal Plants Research 5: 663-676

Zhou ML, Zhu XM, Shao JR, Wu YM, Tang YX (2012) A protocol for genetic transformation of *Catharanthus roseus* by Agrobacterium rhizogenes A4. Applied Biochemistry and Biotechnology 166: 1674-1684

7 Appendixes

7.1 Appendix 1. *Agrobacterium*-mediated transformation

Currently, numerous transformation methods allowing the overexpression of specific genes in plants are available. These can be divided into two main groups: direct and indirect methods. The indirect methods of plant transformation are based on the introduction of a plasmid carried transferred-DNA (T-DNA) delivering the transgene into the target cell by means of bacterial infection, either with *Agrobacterium tumefaciens* or *Agrobacterium rhizogenes,* which are gram negative soil bacteria. The direct methods cause physical or chemical modifications on the cell wall or the plasma membrane, therefore allowing the introduction of the exogenous naked DNA into the cell.

The direct methods avoid the complex interaction between bacterium and plant tissue, with the result that the DNA to be introduced does not require the sequences necessary for T-DNA replication and transfer (Rao et al., 2009). However, *Agrobacterium*-mediated transformation is by large the most efficient method for the introduction of foreign genes into plants, being applied to a growing number of plant species (Tzfira et al., 2007).

Agrobacterium-mediated transformation has been performed over a broad group of organisms from different kingdoms, from plants to fungi. The success of the transformation relies strongly on the selected *Agrobacterium* strain, that may cause neoplastic diseases, crown gall (*Agrobacterium tumefaciens*) or hairy root formation (*Agrobacterium rhizogenes*). The later can be excised and grown *in vitro* as hairy root cultures (Shanks and Morgan, 1999; Gelvin, 2003; Gelvin, 2009).

Genetic experiments indicated that a particular class of plasmids, the tumour-inducing (Ti), and later the rhizogenic (Ri) plasmids, were responsible for the tumorigenesis induced by, respectively, *Agrobacterium tumefaciens* and *Agrobacterium rhizogenes.* Both plasmids, the Ti-plasmid and the Ri-plasmid, share extensive functional homology. A portion of these virulence plasmids, the T-DNA, is transferred to plant cells and incorporated into the plant genome (Gelvin, 2003; Citovsky et al., 2006; Gelvin, 2009).

The T-DNA confers to the plant cells the ability to grow in the absence of exogenous plant hormones. A distinctive biochemical feature of T-DNA transformed plant tissues is its ability to synthetize opines, which are sugar and amino acids conjugates specifically

catabolized by the inciting bacteria, as a C and N source (Vilas P. Sinkar, 1987; Bhadra, 1994). The transfer of the T-DNA to the plant genome is mediated by another segment on the plasmid known as the virulence (vir) region, but the mechanism involved in T-DNA integration in the plant genome has not yet been fully characterized, although it is considered that the integration occurs by illegitimate recombination (Vilas P. Sinkar, 1987). Recombinant disarmed *Agrobacterium* strains, in which the T-DNA oncogenes have been replaced with genes of interests, are today the most efficient vehicles used today for the introduction of foreign genes into plants and for the production of transgenic plant species (Tzfira et al., 2007).

Ri plasmids, and accordingly *Agrobacterium rhizogenes* strains, have been assigned to three classes depending on the opine synthesized and degraded, mannopine, cucumopine, and agropine. In the first two classes, T-DNA consists of a single region, whereas in the third class two T- DNA regions are present, T_L-DNA and T_R-DNA. The two tDNAs of the agropine-type strains, where *Agrobacterium rhizogenes* A4 is included, are separated from each other by about 15 Kb of non-transferred DNA (Bhadra, 1994; Palazon et al., 1998). The right T_R–DNA contains genes homologous to the T-DNA from Ti- plasmids. Most important among these are the *aux* genes (Vilas P. Sinkar, 1987; Palazon et al., 1998). The left T_L-DNA is about 20 Kb in length but, unlike the T_R–DNA does not appear to be closely related to any other characterized Ti-plasmid. Transposon mutagenesis of the T_L-DNA has revealed the presence of at least four loci (*rolA, B, C* and *D*) which affect tumorigenesis in infected plants. Transformed roots arise as a result of the integration of T_L-DNA and T_R-DNA into the plant genome by the agropine strains, and its expression induces root differentiation and subsequent growth (Palazon et al., 1998; Batra et al., 2004; Zhou et al., 2012).

Hairy roots, due to their genetic stability, have received considerable attention from plant biotechnologists for the production of secondary plant compounds. They can be grown on a hormone-free medium and possess the whole biosynthetic potential of wild type roots and, moreover, are able to regenerate whole viable plants and maintain their genetic stability during further sub-culturing and plant regeneration (Palazon et al., 1998; Giri and Narasu, 2000; Batra et al., 2004; Zhou et al., 2012).

The visible characteristics of the hairy-roots syndrome induced by the Ri-plasmid of *A. rhizogenes* include extensive proliferation of the roots from the infected plant wound, modified gravitropic response, reduced apical dominance and increased lateral branching (Bhadra, 1994).

Nowadays, the binary vector system is used almost universally for *Agrobacterium*-mediated transformation (Gelvin, 2003). This strategy involves a separation of the T-DNA region and the *vir* genes into two different replicons. The *vir* gene functions are provided by the disarmed Ri plasmids resident in the *Agrobacterium* strain. The T-DNA, within which are the genes to be transferred, is provided on another replicon, known as binary vector (Hellens et al., 2000). When these replicons are within the same *Agrobacterium* cell, products of the *vir* genes can act in trans on the T-region to effect T-DNA processing and transfer to a plant cell (A. Hoekema, 1983; Framond, 1983). The great advantage of this system is that the binary vector plasmids are small and easy to manipulate in both *E. coli* and *Agrobacterium*, and generally contain multiple unique restriction endonuclease sites within the T-region, into which genes of interest can be cloned (Gelvin, 2003). Therefore, a number of binary vectors were developed such as the pGreen collection, pPVP, pBIN and pCambia (Wydro et al., 2006).

The pGreen vector, selected to be used in this work, is based on the general-purpose cloning vector pBluescript (Alting-Mees and Short, 1989) and therefore contains a pColE1 ori for replication in *E. coli*. The *nptI* gene (encoding kanamycin resistance) from pACYC 177 was also included in the pGreen plasmid (Chang and Cohen, 1978). The pGreen helper plasmid, pSoup carries the pSa replicase gene, *rep* A, that provides replication functions in *trans,* and tetracycline resistance gene (*tet*) from pAlter (Promega) (Roger P. Hellens, 2000). The pSoup offers a substantial saving on the size of the binary vector, pGreen.

The reduced plasmid size, the transformation selection flexibility and the extensive multiple cloning site are the major advantages of pGreen. Versions of pGreen have been used to transform several plant species with the same efficiencies as other binary Ti vectors. Information on the pGreen plasmid system is supplemented by an Internet site (http://www.pgreen.ac.uk) (Roger P. Hellens, 2000).

7.2 Appendix 2. Modificated Hanahan's protocol for the preparation of *E. coli* DH5α chemically competent cells

Procedure

- Inoculate a single colony from a Luria-Berthani (LB) (Difco TM) with agar (1.5% (w/v), Liofilchem) plate in 25 mL of LB medium. Incubate overnight at 37°C with vigorous shaking (200 rpm).
- Dilute the incubated culture in 225 mL (~1:100) of LB medium supplemented with 10 mM of MgCl 2 and 10 mM of MgSO 4. Incubate at 37°C with vigorous shaking until the culture reaches an $A_{600nm} \approx 0.6$.
- Incubate the cells for 10 min on ice. Centrifuge at 1500 g (3000 rpm) for 5 min at 4°C to pellet the cells.
- Pour off supernatant. Remove the remaining supernatant by pipetting without disturbing the pellet. Resuspend the pelleted cells in 100 mL of chilled RF1.
- Incubate the resuspended cells for 15 min on ice. Centrifuge at 1500 g (3000 rpm) for 5 min at 4°C.
- Discard supernatant as described before. Resuspend the pelleted cells in 16 mL of chilled RF2.
- Incubate on ice for 45 min to 1 h. Prepare 50 or 100 µL aliquots into prechilled 1.5 mL eppendorf tubes. Flash freeze the tubes in liquid N_2 and store at -80°C.

Solutions

RF1

Composition	**per 250mL**
30 mM potassium acetate	0.74 g
10 mM CaCl 2	0.37 g
50 mM MnCl 2	2.47 g
100 mM RbCl	3.02 g
15% (v/v) glycerol	37.5 mL

Adjust pH to 5.8 with 0.1M NaOH.
Filter-sterilize (0.2 µm).

RF2

Composition	**per 100 mL**
10 mM MOPS	0.335 g
75 mM CaCl 2	3.75 mL of a 2 M stock solution
10 mM RbCl	0.12 g
15% (v/v) glycerol	15 mL

Adjust pH to 8.0 with 1M NaOH.
Filter-sterilize (0.2 µm).

LB medium

Composition	**per 1 L**
1% (w/v) tryptone	10 g
1% (w/v) NaCl	10 g
0.5% (w/v) yeast extract	5 g

Sterilize by autoclaving.

7.3 Appendix 3. Modificated Hanahan's protocol for the preparation of *E. coli TOP10* chemically competent cells.

Procedure

- Inoculate a single colony from a LB plate in 25 mL of LB medium. Incubate overnight at 37°C with vigorous shaking (200 rpm).
- Dilute the incubated culture in 225 mL (~1:100) of LB medium supplemented with 10 mM of MgCl 2 and 10 mM of MgSO 4. Incubate at 37°C with vigorous shaking until the culture reaches an A600nm ≈ 0.5.
- Incubate the cells for 10 min on ice. Centrifuge at 2300 g (5000 rpm) for 10 min at 4°C to pellet the cells.
- Pour off supernatant. Remove the remaining supernatant by pipetting without disturbing the pellet. Resuspend the pelleted cells in 80 mL of chilled CCMB80 buffer.
- Incubate the resuspended cells for 20 min on ice. Centrifuge at 2300 g (5000 rpm) for 10 min at 4°C.
- Discard supernatant as described before. Resuspend the pelleted cells in 10 mL of chilled CCMB80 buffer.
- Test OD of a mixture of 200 µl SOC and 50 µl of the resuspended cells. Add ice cold CCMB80 buffer to the whole suspension to yield a final OD of 1.0-1.5 in this test.
- Incubate on ice for 45 min to 1 h. Prepare 50 or 100 µL aliquots into prechilled 1.5 mL tubes. Flash freeze the tubes in liquid N_2 and store at -80°C.

Solutions

CCMB80 buffer

Composition	**per 1 L**
10 mM potassium acetate	10 mL of a 1 M stock solution
80 mM CaCl 2.2H 2O	11.8 g
20 mM MnCl 2.4H 2O	4.0 g
10 mM MgCl 2.6H 2O	2.0 g
10% (v/v) glycerol	100 mL

Adjust pH down to 6.4 with 0.1N HCl.
Filter-sterilize (0.2 µm).

7.4 Appendix 4. Protocol for the preparation of *A. rhizogenes A4* and *A. tumefaciens C58C1* chemically competent cells

Procedure

- Inoculate a single colony from a YEB with agar (1.8% (w/v), Liofilchem) plate in 5 mL of YEB medium. Incubate overnight at 28°C with vigorous shaking (200 rpm).
- Dilute 2mL of the incubated culture in 50 mL (1:25) of YEB medium. Incubate at 28°C with vigorous shaking until the culture reaches an $A_{600nm} \approx 0.5\text{-}1.0$.
- Cooldown the culture on ice. Centrifuge at 3000 g for 5 min at 4°C to pellet the cells.
- Pour off supernatant. Remove the remaining supernatant by pipetting without disturbing the pellet. Resuspend the pelleted cells in 1 mL of chilled $CaCl_2$ 20mM.
- Prepare 50 or 100 µL aliquots into prechilled 1.5 mL eppendorf tubes. Flash freeze the tubes in liquid N_2 and store at -80°C.

YEB medium

Composition	**per 1 L**
0.5% (w/v) beef extract	5 g
0.1% (w/v) yeast extract	1 g
0.5% (w/v) peptone	5 g
0.5% (w/v) sucrose	5 g

Adjust pH to 7.2. Sterilize by autoclaving.

7.5 Appendix 5. DNA extraction protocol with CTAB

Procedure

- Use a mortar and pestle to homogenize the plant tissue (hairy roots) in liquid nitrogen.
- Transfer to an eppendorf and add 1 mL CTAB (pre-heated at 65°C). Invert several times.
- Incubate at 65°C for 30 min – heating block- (gently reverse the tube before the incubation and at the 15 min).
- Cool down the eppendorf (HOTE).
- Add 500 µL chloroform:isoamyl alcohol (24:1). Shake and centrifuge at 5900 g (8000 rpm) for 5 min. Transfer the upper phase to a new eppendorf. Throw away the lower phase.
- Newly, add 500 µL chloroform:isoamyl alcohol (24:1). Shake and centrifuge at 5900 g (8000 rpm) for 5 min. Transfer the upper phase to a new eppendorf. Throw away the lower phase.
- Add 500 µL of cold isopropanol, mix gently and leave by the bench for 20 min.
- Centrifuge at 15700 g (13000 rpm) for 5 min. Throw away the supernatant and air dry the pellet – stove 55°C.
- Ressuspend the pellet in 400 µL of water and add 1 µL of 10 $mg.mL^{-1}$ of RNase and incubate at 37°C for 30 min.
- Add 400 µL chloroform:isoamyl alcohol (24:1). Shake and centrifuge at 5900 g (8000 rpm) for 5 min. Transfer the upper phase to a new eppendorf. Throw away the lower phase.
- Add 2.5 vol of cold ethanol 100%. Leave at -20°C for 10 min.
- Centrifuge at 15700 g (13000 rpm) for 5 min. Throw away the supernatant.
- Add 500 µL of cold ethanol 70% and centrifuge at 15700 g (13000 rpm) for 5 min.
- Throw away completely the supernatant and dry the pellet at 55°C for 10 min.
- Resuspend the DNA in 50 µL sterile water. Leave at 4°C for 30 min to complete resuspension. Do not shake!
- Store at 4°C.

Solutions

CTAB

Composition	**per 100 mL**
Tris 1M pH 8.0	10 mL
NaCl 5M	28 mL
EDTA 0.5M	4 ml
CTAB	2 g
b-mercaptoethanol	10 µl
PVP	0.08 gr

Sterilize by autoclaving.

7.6 Appendix 6. Results of the BLAST search for the retrieval of the full coding sequences of the transcription factor genes.

CrTF12

Tag and MPGR best hit alignment

```
>cra_locus_12026_iso_4_len_5472_ver_3
      Length = 5472

 Minus Strand HSPs:

 Score = 1235 (191.3 bits), Expect = 6.6e-50, P = 6.6e-50
 Identities = 251/253 (99%), Positives = 251/253 (99%), Strand = Minus / Plus

Query:    252 TTAAAATGGCTTCTGCAAGA-CATGAACATATCCTGTGAGAGGCGACATCTGATAATTGC 194
              |||||||||||||||||| |||||||||||||| ||||||||||||||||||||||
Sbjct:    363 TTAAAATGGCTTCTGCAAGAACATGAACATATCCTG-GAGAGGCGACATCTGATAATTGC 421

Query:                                                                193
TCGGAAACTGAGTGTTCATAAGCACATTGCCGTCGTAACTGAATGGAAAATCAAGGGCAG 134
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:    422 TCGGAAACTGAGTGTTCATAAGCACATTGCCGTCGTAACTGAATGGAAAATCAAGGGCAG 481

Query:    133 AAGTGGATTTCTCCCAGTCATTCAGCTTCGGCTCGCTCTGTACTTCGCATGTGTTCTGAG 74
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:    482 AAGTGGATTTCTCCCAGTCATTCAGCTTCGGCTCGCTCTGTACTTCGCATGTGTTCTGAG 541

Query:                                                                 73
AGAATTCCGGCGACACCACTTGCTCAGATGAGCTGGAGTTCGTGTGCAGCTTGGGAACGG 14
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:                                                                542
AGAATTCCGGCGACACCACTTGCTCAGATGAGCTGGAGTTCGTGTGCAGCTTGGGAACGG 601

Query:    13 AATCTGATGGATC 1
              |||||||||||||
Sbjct:   602 AATCTGATGGATC 614
```

MPGR - Retrieved protein sequence

```
>cra_locus_12026_iso_4_len_5472_ver_3
MAAELQLPPGFRFHPTDEELVIHYLCRKCASQPIAVPIIAEIDLYKYNPWDLPGMALYGEKEWY
FFSPRDRKYPNGSRPNRAAGSGYWKATGADKPIGNPKPMGIKKALVFYAGKAPKGEKTNWIMHE
YRLADVDRSARKRNNSLRLDDWVLCRIYNKKGTIDKQQMMGIRKAASQVIEEEDKKPEILKSVP
ETTPVVYSDFMYLDPSDSVPKLHTNSSSSEQVVSPEFSQNTCEVQSEPKLNDWEKSTSALDFPF
SYDGNVLMNTQFPSNYQMSPLQDMFMFLQKPF-
```

MPGR- Retrieved nucleotide sequence

>cra_locus_12026_iso_4_len_5472_ver_3
CCAGCCGCCTCTCCTTGCCACGTGTCCCTATTCTACCCTTATATATTTCCCATCTTCCTCTATG
TCTGAAGTTTCAACCTTCAAGCCCCCAATTTCCCCATAGCTCTCTACCTCCTCCCCCTCTCTATT
TTTTCTCTCTAGAGTAGACTTTTCTAGTGAGAAGAAAACAAAAAAAAGAAGAAAAAATGGCTGC
AGAGTTGCAATTACCGCCTGGATTCAGATTTCATCCGACCGATGAAGAACTCGTGATACATTAT
TTATGCCGGAAATGTGCGTCTCAACCGATTGCGGTCCCGATTATAGCTGAGATTGATCTCTATA
AGTATAATCCTTGGGATCTTCCTGGTATGGCGTTGTACGGGGAGAAGGAGTGGTATTTTTTTTC
TCCGAGGGACAGAAAGTATCCGAACGGTTCGAGGCCGAATAGGGCGGCCGGAAGTGGTTATTGG
AAGGCGACCGGTGCGGATAAACCGATTGGAAACCCGAAACCGATGGGGATTAAGAAAGCCTTGG
TGTTTTACGCCGGCAAAGCCCCAAAAGGAGAGAAGACTAATTGGATTATGCATGAATACAGACT
CGCCGATGTTGATCGATCCGCCCGTAAGAGGAACAACAGCTTAAGGTTGGATGATTGGGTTCTG
TGCCGCATATACAACAAGAAAGGTACAATCGACAAACAACAAATGATGGGCATTCGTAAAGCGG
CAAGTCAGGTGATTGAAGAAGAGGATAAGAAGCCGGAAATTCTGAAGTCGGTGCCGGAAACAAC
ACCGGTGGTGTACAGCGATTTCATGTACCTGGATCCATCAGATTCCGTTCCCAAGCTGCACACG
AACTCCAGCTCATCTGAGCAAGTGGTGTCGCCGGAATTCTCTCAGAACACATGCGAAGTACAGA
GCGAGCCGAAGCTGAATGACTGGGAGAAATCCACTTCTGCCCTTGATTTTCCATTCAGTTACGA
CGGCAATGTGCTTATGAACACTCAGTTTCCGAGCAATTATCAGATGTCGCCTCTCCAGGATATG
TTCATGTTCTTGCAGAAGCCATTTTAAAAGACAAGAGGGAGGATGAAAATTCCTGGAAATTTTT
ACATTTTCTGCTCGCACGTGTAACATGAGACTTAGGTTTATGAGTGTACACGTGCGAACTCATT
GGCATAGGCAGTAAAATATGGGTATTAGAATTGCTTGGGCAGGCAATTGTTAATTAATACATAG
TATTATTACTTTTCATTAATTATTGTGATGTATATAATATATATATATATATTTTACTCAGTTG
ATGATACCATTTTGAGGGAAATGGTAAATGTTAGTGTAGAGAGAGGGTATATTAGTCAAATTAT
ATGTTCCAATCTATGAACTTAGCCTCCATTAAAAGCATTAATAATTCAGTCAGTCAAGAGAGAT
CATCCTTACGACGGCAATGTGCTTATGAACACTCAGTTTCCGAGCAATTATCAGATGTCGCCTC
TCCAGGATATGTTCATGTTCTTGCAGAAGCCATTTTAA

Start Codon

Stop Codon

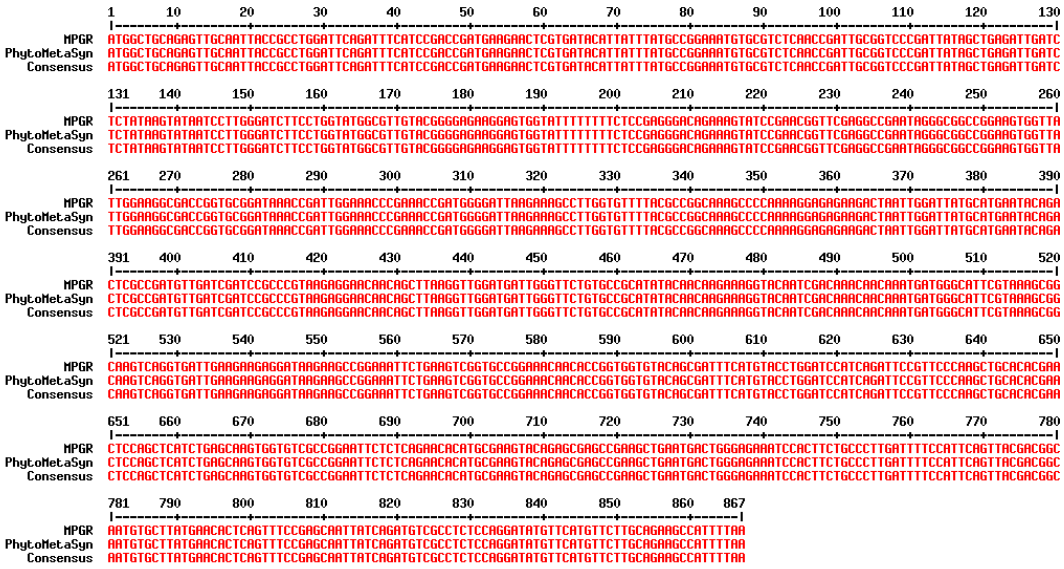
Coding Sequence

PhytoMetaSyn - Retrieved nucleotide sequence

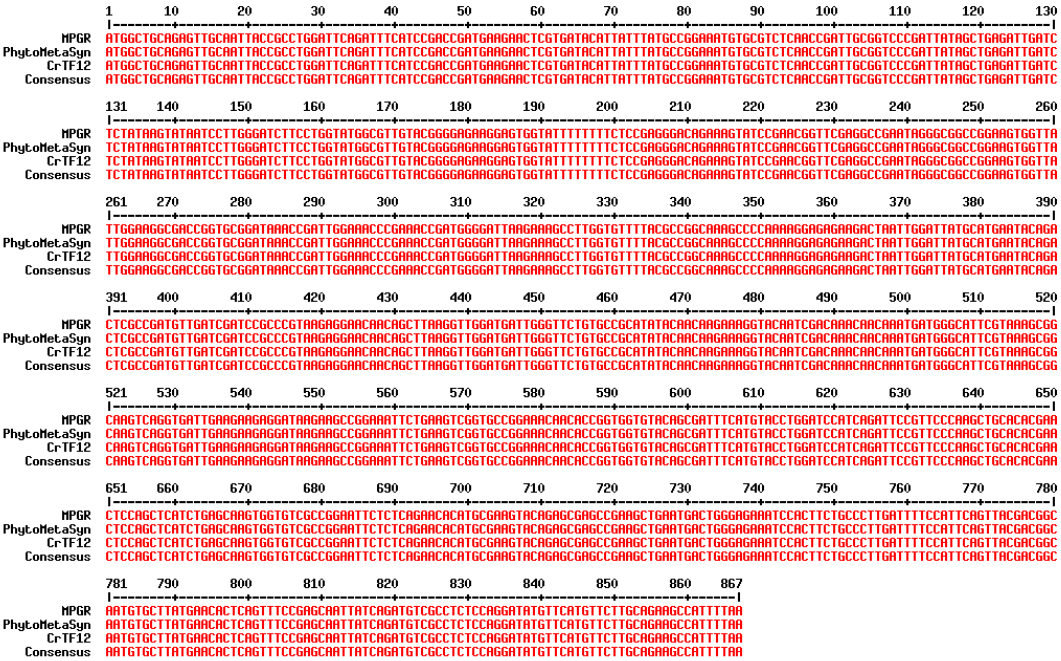
>gnl|MAGPIE|cro.CROWL1VD_mira--CROWL1VD_rep_c1072 No definition line found
CTACCTCCTCCCCTCTCTATTTTTTCTCTCTAGAGTAGACTTTTCTAGTGAGAAGAAAACAAAA
AAAGAAGAAAAAATGGCTGCAGAGTTGCAATTACCGCCTGGATTCAGATTTCATCCGACCGATG
AAGAACTCGTGATACATTATTTATGCCGGAAATGTGCGTCTCAACCGATTGCGGTCCCGATTAT
AGCTGAGATTGATCTCTATAAGTATAATCCTTGGGATCTTCCTGGTATGGCGTTGTACGGGGAG
AAGGAGTGGTATTTTTTTTCTCCGAGGGACAGAAAGTATCCGAACGGTTCGAGGCCGAATAGGG
CGGCCGGAAGTGGTTATTGGAAGGCGACCGGTGCGGATAAACCGATTGGAAACCCGAAACCGAT
GGGGATTAAGAAAGCCTTGGTGTTTTACGCCGGCAAAGCCCCAAAAGGAGAGAAGACTAATTGG
ATTATGCATGAATACAGACTCGCCGATGTTGATCGATCCGCCCGTAAGAGGAACAACAGCTTAA
GGTTGGATGATTGGGTTCTGTGCCGCATATACAACAAGAAAGGTACAATCGACAAACAACAAAT
GATGGGCATTCGTAAAGCGGCAAGTCAGGTGATTGAAGAAGAGGATAAGAAGCCGGAAATTCTG
AAGTCGGTGCCGGAAACAACACCGGTGGTGTACAGCGATTTCATGTACCTGGATCCATCAGATT
CCGTTCCCAAGCTGCACACGAACTCCAGCTCATCTGAGCAAGTGGTGTCGCCGGAATTCTCTCA
GAACACATGCGAAGTACAGAGCGAGCCGAAGCTGAATGACTGGGAGAAATCCACTTCTGCCCTT

GATTTTCCATTCAGTTACGACGGCAATGTGCTTATGAACACTCAGTTTCCGAGCAATTATCAGA
TGTCGCCTCTCCAGGATATGTTCATGTTCTTGCAGAAGCCATTTTAAAAGACAAGAGGGAGGAT
GAAAATTCCTGGAAATTTTTACATTTTCTGCTCGCACGTGTAACATGAGACTTAGGTTTATGAG
TGTACACGTGCGAACTCATTGGCATAGGCAGTAAAATATGGGTATTAGAATTGCTTGGGCAGGC
AATTGTTAATTAATACATAGTATTATTACTTTTCATTAATTATTGTGATGTATATAATATATAT
ATATATATTTTACTCAGTTGATGATACCATTTTGAGGGAAATGGTAAATGTTAGTGTAGAGAGA
GGGTATATTAGTCAAATTATATGTTCCAATCTATGAACTTAGCCTCCATTAAAAGCATTAATAA
TTCAGTCAGTCAAGAGAGATCAKCCAATTAAAAATGGTTTAATTATATTTTTTATAAATTTTAA
ATACATTATTATTTTATT

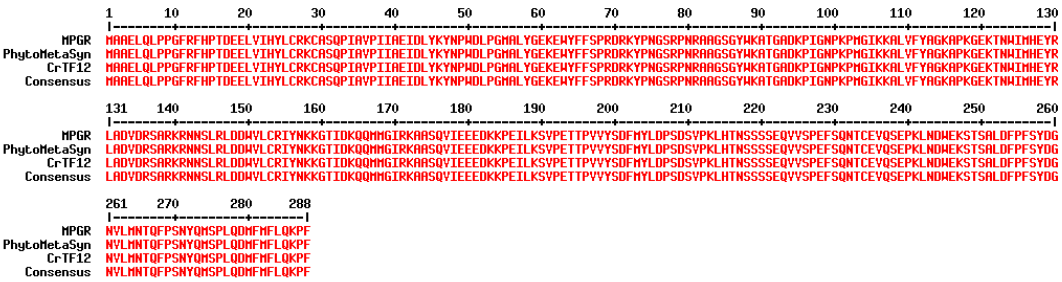
Alignment of the retrieved nucleotide sequences from MPGR and PhytoMetaSyn databases



Alignment of the retrieved nucleotide sequences from MPGR and PhytoMetaSyn databases, and from sequenced *CrTF12*



Alignment between the proteins translated from the nucleotide sequences retrieved from MPGR and PhytoMetaSyn databases, and from sequenced *CrTF12*



CrTF19

Tag and MPGR best hit alignment

>cra_locus_517_iso_2_len_3386_ver_3
Length = 3386

Plus Strand HSPs:

Score = 1446 (223.0 bits), Expect = 8.3e-60, P = 8.3e-60
Identities = 296/300 (98%), Positives = 296/300 (98%), Strand = Plus / Plus

```
Query:                                                                    1
GATCCAGTTGATGCATGGTGGAGTGATGGCTGGTGGGAGGGAGTACTAACTGCCTCCAAC 60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:                                                                 2171
GATCCAGTTGATGCATGGTGGAGTGATGGCTGGTGGGAGGGAGTACTAACTGCCTCCAAC 2230

Query:       61  AACTCAGGAAGTGAAAATTATCAAGTTTATATTTCCAGTGAAAACTTTTTTTTGACCATA
120
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:     2231  AACTCAGGAAGTGAAAATTATCAAGTTTATATTTCCAGTGAAAACTTTTTTTTGACCATA
2290

Query:                                                                  121
GACAAAAAGCTGCTAAGGGCTTCAAGAGATTGGGTTGGAGGTCAATGGGTTGATGTGGAG 180
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:                                                                 2291
GACAAAAAGCTGCTAAGGGCTTCAAGAGATTGGGTTGGAGGTCAATGGGTTGATGTGGAG 2350

Query:      181  CGAAATCCTGAAGTTCTTTCAAATCTTAACTGCAGCTTTGAGTTTAGAG--TAAGCTTTC
238
              |||||||||||||||||| ||||| ||||||||||||||||||  ||||||||
Sbjct:     2351  CGAAATCCTGAAGTTCTTTCAA-TCTTATCTGCAGCTTTGAGTTTAGAGAGTAAGCTTTC
2409

Query:   239 TGTATCCTCGGCTTTCTCCAAGGAGGCGAAGTCTGATGGTTCTCCTATGTCATGCCTTAA
298
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  2410 TGTATCCTCGGCTTTCTCCAAGGAGGCGAAGTCTGATGGTTCTCCTATGTCATGCCTTAA
2469
```

MPGR - Retrieved protein sequence

```
MEGTICFEGEGKPYCSLFFEGFIRANRIVRYFLKDSSGDSVLAVVGTERSVRHMFYVIAEEFLN
AYGAENSVHAGFRWRSRREVVNWLTSMLSKQHRQSDCSRSPNDDLSFGVNVQMQMQEPKGRLAR
NLRGHPSDIVWSGVAWTCGKQLKHYPAFCRNGITIATQSFVYVMAEKENRHLAYLEDMYEDRKS
QKKVKVRWFHHNREVKGVVALRNPHPKEVFITPYAQVISAECVDGPAIVLTREHYEKCVSVFPN
DLLARAHFCFRQFKSNRVKPFKLSKLCGYFDQQIFSCFSPDFFEDEQFSSGDEVKVGVKRTRSC
RERAKVTNKLSYQKLRYSMLSRRLNCHKNERQSWQIPIFSVSEKVEFLCQDSGLRGCWFRCTVL
EVSRRQIKIQYDDIKDEDGCGNLEEWVPAFRLANPDKFGMRYSGRSTLRPAVTSLPSEVAFKVG
DPVDAWWSDGWWEGVLTASNNSGSENYQVYISSENFFLTIDKKLLRASRDWVGGQWVDVERNPE
VLSILSAALSLESKLSVSSAFSKEAKSDGSPMSCLKVPSDIKLSVVYETETQQKDWPNGNQKPV
VKDEDAENTAKHDDNDLKINGDSSDQKNVNVESLLTDNNLVDKSDNSDEDNKSAELEAARQEC
```

MPGR- Retrieved nucleotide sequence

```
>cra_locus_517_iso_2_len_3386_ver_3
GGTTCTTGTAAAGTTTGCAGCTTTTGGGCTGCAAAAGTTTGAGTTGGGAAACACTTTTTGGACG
TCGCTTTGTTTGGGCATCTTTAGCTTTACAGATTAGAGTTCTTTTAGGGATTTGATGGTTCATT
CTGCCCCTTAGGTCGCCGTGGTTTTCTACCTAAAATGTCTGCAAATGATCACTGCTTTGTGGA
ATGGAAGGAACAATTTGTTTCGAAGGAGAGGGGAAACCGTATTGTTCACTATTTTTTGAAGGAT
TCATCAGGGCAAACCGCATTGTTCGCTATTTTTTGAAGGATTCATCAGGGGACTCTGTCCTTGC
TGTTGTGGGTACAGAGAGAAGTGTTAGGCACATGTTCTATGTTATTGCCGAGGAGTTTCTGAAT
GCATATGGAGCTGAGAACTCTGTCCATGCTGGCTTCAGATGGAGATCAAGGAGAGAGGTCGTCA
ATTGGCTTACTTCTATGCTATCCAAGCAGCATAGACAAAGTGATTGCTCGAGATCACCAAACGA
TGACTTAAGCTTTGGAGTTAATGTTCAAATGCAGATGCAAGAACCTAAGGGTCGTCTTGCAAGG
AATCTCAGAGGACATCCTTCAGATATAGTATGGTCAGGTGTAGCATGGACATGTGGTAAACAGC
```

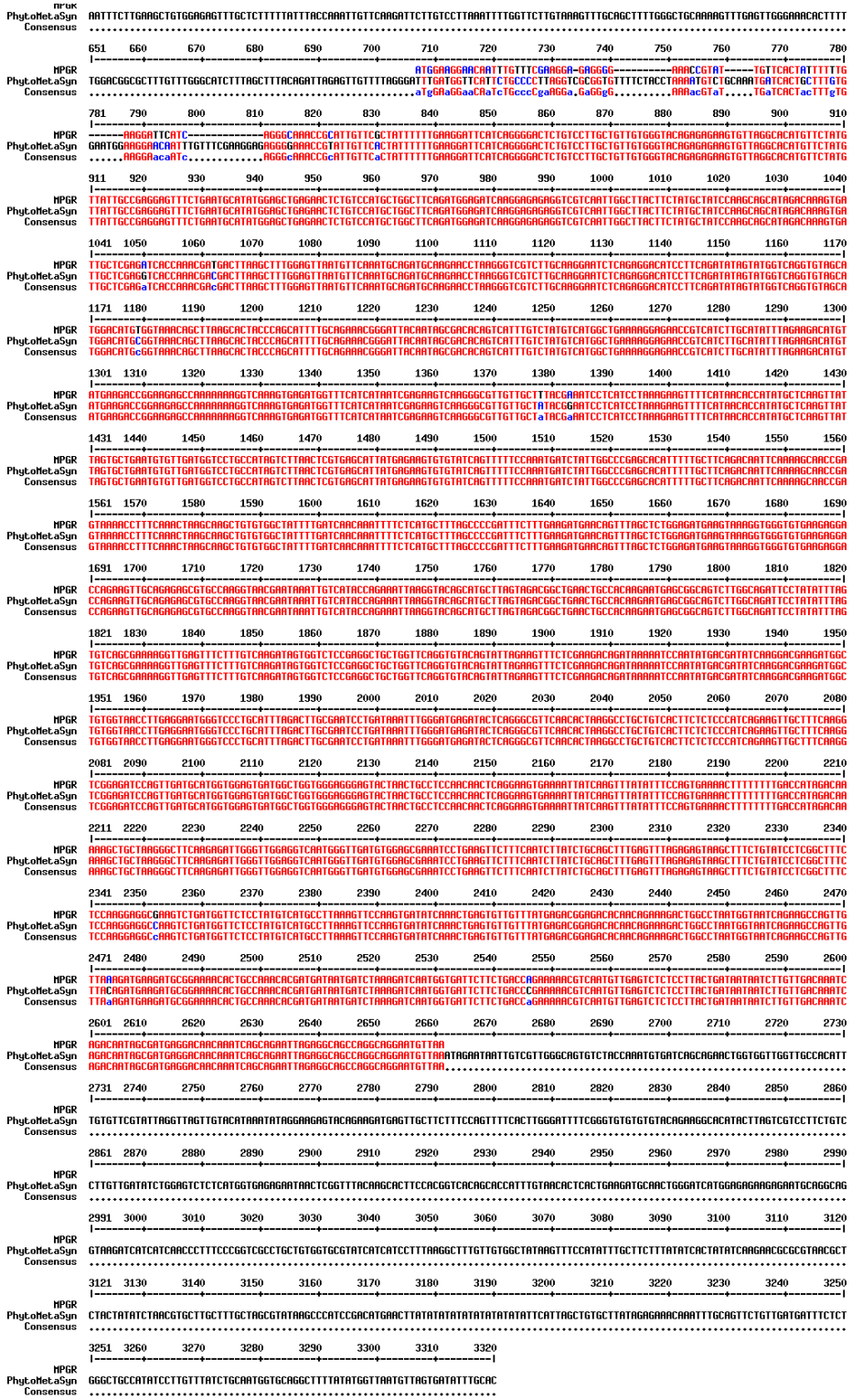
TTAAGCACTACCCAGCATTTTGCAGAAACGGGATTACAATAGCGACACAGTCATTTGTCTATGT
CATGGCTGAAAAGGAGAACCGTCATCTTGCATATTTAGAAGACATGTATGAAGACCGGAAGAGC
CAAAAAAAGGTCAAAGTGAGATGGTTTCATCATAATCGAGAAGTCAAGGGCGTTGTTGCTTTAC
GAAATCCTCATCCTAAAGAAGTTTTCATAACACCATATGCTCAAGTTATTAGTGCTGAATGTGT
TGATGGTCCTGCCATAGTCTTAACTCGTGAGCATTATGAGAAGTGTGTATCAGTTTTTCCAAAT
GATCTATTGGCCCGAGCACATTTTTGCTTCAGACAATTCAAAAGCAACCGAGTAAAACCTTTCA
AACTAAGCAAGCTGTGTGGCTATTTTGATCAACAAATTTTCTCATGCTTTAGCCCCGATTTCTT
TGAAGATGAACAGTTTAGCTCTGGAGATGAAGTAAAGGTGGGTGTGAAGAGGACCAGAAGTTGC
AGAGAGCGTGCCAAGGTAACGAATAAATTGTCATACCAGAAATTAAGGTACAGCATGCTTAGTA
GACGGCTGAACTGCCACAAGAATGAGCGGCAGTCTTGGCAGATTCCTATATTTAGTGTCAGCGA
AAAGGTTGAGTTTCTTTGTCAAGATAGTGGTCTCCGAGGCTGCTGGTTCAGGTGTACAGTATTA
GAAGTTTCTCGAAGACAGATAAAAATCCAATATGACGATATCAAGGACGAAGATGGCTGTGGTA
ACCTTGAGGAATGGGTCCCTGCATTTAGACTTGCGAATCCTGATAAATTTGGGATGAGATACTC
AGGGCGTTCAACACTAAGGCCTGCTGTCACTTCTCTCCCATCAGAAGTTGCTTTCAAGGTCGGA
GATCCAGTTGATGCATGGTGGAGTGATGGCTGGTGGGAGGGAGTACTAACTGCCTCCAACAACT
CAGGAAGTGAAAATTATCAAGTTTATATTTCCAGTGAAAACTTTTTTTTGACCATAGACAAAAA
GCTGCTAAGGGCTTCAAGAGATTGGGTTGGAGGTCAATGGGTTGATGTGGAGCGAAATCCTGAA
GTTCTTTCAATCTTATCTGCAGCTTTGAGTTTAGAGAGTAAGCTTTCTGTATCCTCGGCTTTCT
CCAAGGAGGCGAAGTCTGATGGTTCTCCTATGTCATGCCTTAAAGTTCCAAGTGATATCAAACT
GAGTGTTGTTTATGAGACGGAGACACAACAGAAAGACTGGCCTAATGGTAATCAGAAGCCAGTT
GTTAAAGATGAAGATGCGGAAAACACTGCCAAACACGATGATAATGATCTAAAGATCAATGGTG
ATTCTTCTGACCAGAAAAACGTCAATGTTGAGTCTCTCCTTACTGATAATAATCTTGTTGACAA
ATCAGACAATAGCGATGAGGACAACAAATCAGCAGAATTAGAGGCAGCCAGGCAGGAATGTTAA
ATAGAATAATTGTCGTTGGGCAGTGTCTACCAAATGTGATCAGCA

PhytoMetaSyn - Retrieved nucleotide sequence

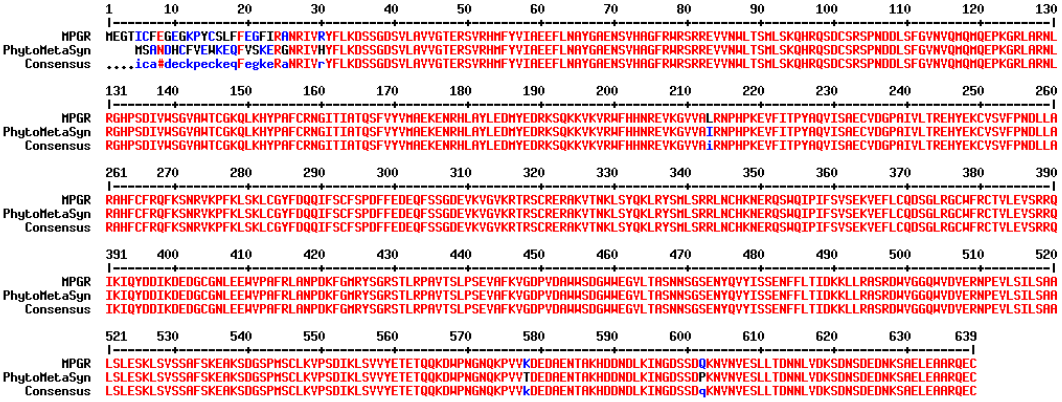
>gnl|MAGPIE|cro.CROAJ1VD_velvet--Singlet5841 No definition line found
CGCATCCGCCACCACACAACCAACCTTAATCCCCAAAACCTTCTCCTCTTTAGCCTCTGCCATG
TTTGCTCTTTCTATTCACTCTCTTGCCACCAATCAACACATTTACCATTTCCCCACCTCATCTT
GTCAACATCAACTATAGTATGCTGCCTCTTCCATAGCTGATCTTTACAGAGAAATAGAAGAAAT
TAGCTTCTCTGTCTCCGTCTTTCTTGGTTCTACTGCATAGTCAGACTCAGGCCTCTGTTCAGCT
CCGGAGCTGTTTATAATTCTTTTTGTGTGTTGTTTGTAATGGGTTACTGCAAATTCGACTGCCC
ATTTCGATATTTTTTGGAATTTTGCTATCTCAGTGTGACTTCTTGTTGTTAGGTGGGCTATGTG
TGAGTTTAATCTATCTTCTTTCTTGGGTTTTGATCTTCTTCACGGATCTTTGTCACCACATGGT
TTTGAGGGATATATTTAGTTTCGGTAAAAAAAAACCCAAGTCGTCTGTTAAATGTTCTTTGACT
TTAGGCTGAATTTCTTGAAGCTGTGGAGAGTTTGCTCTTTTTATTTACCAAATTGTTCAAGATT
CTTGTCCTTAAATTTTGGTTCTTGTAAAGTTTGCAGCTTTTGGGCTGCAAAAGTTTGAGTTGGG
AAACACTTTTTGGACGGCGCTTTGTTTGGGCATCTTTAGCTTTACAGATTAGAGTTGTTTTAGG
GATTTGATGGTTCATTCTGCCCCTTAGGTCGCGGTGTTTTCTACCTAAAATGTCTGCAAATGAT
CACTGCTTTGTGGAATGGAAGGAACAATTTGTTTCGAAGGAGAGGGGAAACCGTATTGTTCACT
ATTTTTTGAAGGATTCATCAGGGGACTCTGTCCTTGCTGTTGTGGGTACAGAGAGAAGTGTTAG
GCACATGTTCTATGTTATTGCCGAGGAGTTTCTGAATGCATATGGAGCTGAGAACTCTGTCCAT
GCTGGCTTCAGATGGAGATCAAGGAGAGAGGTCGTCAATTGGCTTACTTCTATGCTATCCAAGC
AGCATAGACAAAGTGATTGCTCGAGGTCACCAAACGACGACTTAAGCTTTGGAGTTAATGTTCA
AATGCAGATGCAAGAACCTAAGGGTCGTCTTGCAAGGAATCTCAGAGGACATCCTTCAGATATA
GTATGGTCAGGTGTAGCATGGACATGCGGTAAACAGCTTAAGCACTACCCAGCATTTTGCAGAA

ACGGGATTACAATAGCGACACAGTCATTTGTCTATGTCATGGCTGAAAAGGAGAACCGTCATCT
TGCATATTTAGAAGACATGTATGAAGACCGGAAGAGCCAAAAAAAGGTCAAAGTGAGATGGTTT
CATCATAATCGAGAAGTCAAGGGCGTTGTTGCTATACGGAATCCTCATCCTAAAGAAGTTTTCA
TAACACCATATGCTCAAGTTATTAGTGCTGAATGTGTTGATGGTCCTGCCATAGTCTTAACTCG
TGAGCATTATGAGAAGTGTGTATCAGTTTTTCCAAATGATCTATTGGCCCGAGCACATTTTTGC
TTCAGACAATTCAAAAGCAACCGAGTAAAACCTTTCAAACTAAGCAAGCTGTGTGGCTATTTTG
ATCAACAAATTTTCTCATGCTTTAGCCCCGATTTCTTTGAAGATGAACAGTTTAGCTCTGGAGA
TGAAGTAAAGGTGGGTGTGAAGAGGACCAGAAGTTGCAGAGAGCGTGCCAAGGTAACGAATAAA
TTGTCATACCAGAAATTAAGGTACAGCATGCTTAGTAGACGGCTGAACTGCCACAAGAATGAGC
GGCAGTCTTGGCAGATTCCTATATTTAGTGTCAGCGAAAAGGTTGAGTTTCTTTGTCAAGATAG
TGGTCTCCGAGGCTGCTGGTTCAGGTGTACAGTATTAGAAGTTTCTCGAAGACAGATAAAAATC
CAATATGACGATATCAAGGACGAAGATGGCTGTGGTAACCTTGAGGAATGGGTCCCTGCATTTA
GACTTGCGAATCCTGATAAATTTGGGATGAGATACTCAGGGCGTTCAACACTAAGGCCTGCTGT
CACTTCTCTCCCATCAGAAGTTGCTTTCAAGGTCGGAGATCCAGTTGATGCATGGTGGAGTGAT
GGCTGGTGGGAGGGAGTACTAACTGCCTCCAACAACTCAGGAAGTGAAAATTATCAAGTTTATA
TTTCCAGTGAAAACTTTTTTTTGACCATAGACAAAAAGCTGCTAAGGGCTTCAAGAGATTGGGT
TGGAGGTCAATGGGTTGATGTGGAGCGAAATCCTGAAGTTCTTTCAATCTTATCTGCAGCTTTG
AGTTTAGAGAGTAAGCTTTCTGTATCCTCGGCTTTCTCCAAGGAGGCCAAGTCTGATGGTTCTC
CTATGTCATGCCTTAAAGTTCCAAGTGATATCAAACTGAGTGTTGTTTATGAGACGGAGACACA
ACAGAAAGACTGGCCTAATGGTAATCAGAAGCCAGTTGTTACAGATGAAGATGCGGAAAACACT
GCCAAACACGATGATAATGATCTAAAGATCAATGGTGATTCTTCTGACCCGAAAAACGTCAATG
TTGAGTCTCTCCTTACTGATAATAATCTTGTTGACAAATCAGACAATAGCGATGAGGACAACAA
ATCAGCAGAATTAGAGGCAGCCAGGCAGGAATGTTAAATAGAATAATTGTCGTTGGGCAGTGTC
TACCAAATGTGATCAGCAGAACTGGTGGTTGGTTGCCACATTTGTGTTCGTATTAGGTTAGTTG
TACATAAATATAGGAAGAGTACAGAAGATGAGTTGCTTCTTTCCAGTTTTCACTTGGGATTTTC
GGGTGTGTGTGTACAGAAGGCACATACTTAGTCGTCCTTCTGTCCTTGTTGATATCTGGAGTCT
CTCATGGTGAGAGAATAACTCGGTTTACAAGCACTTCCACGGTCACAGCACCATTTGTAACACT
CACTGAAGATGCAACTGGGATCATGGAGAGAAGAGAATGCAGGCAGGTAAGATCATCATCAACC
CTTTCCCGGTCGCCTGCTGTGGTGCGTATCATCATCCTTTAAGGCTTTGTTGTGGCTATAAGTT
TCCATATTTGCTTCTTTATATCACTATATCAAGAACGCGCGTAACGCTCTACTATATCTAACGT
GCTTGCTTTGCTAGCGTATAAGCCCATCCGACATGAACTTATATATATATATATATATATTCAT
TAGCTGTGCTTATAGAGAAACAAATTTGCAGTTCTGTTGATGATTTCTCTGGGCTGCCATATCC
TTGTTTATCTGCAATGGTGCAGGCTTTTATATGGTTAATGTTAGTGATATTTGCAC

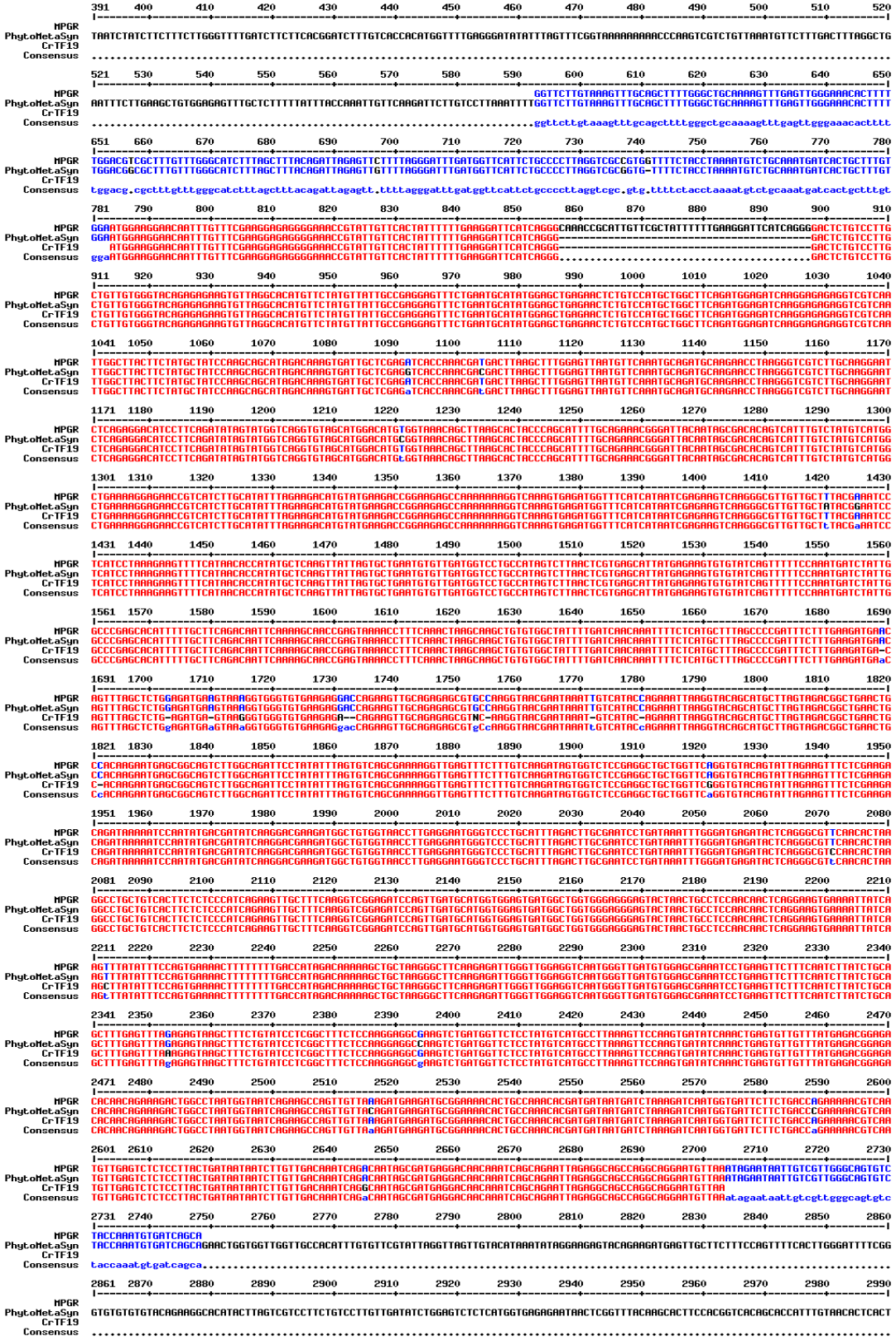
Alignment of the retrieved nucleotide sequences from MPGR and PhytoMetaSyn databases



Alignment between the proteins translated from the nucleotide sequences retrieved from MPGR and PhytoMetaSyn databases



Alignment between the nucleotide sequences retrieved from MPGR and PhytoMetaSyn databases, and from sequenced *CrTF19*



CrTF79

Tag and MPGR best hit alignment

```
>cra_locus_5366_iso_3_len_1889_ver_3
      Length = 1889
 Minus Strand HSPs:
 Score = 1516 (233.5 bits), Expect = 7.7e-64, P = 7.7e-64
 Identities = 322/344 (93%), Positives = 322/344 (93%), Strand = Minus / Plus
Query:   344 AGTTTTCGCCATGTGGACGATGACATGGTCATCCCATTCCATCCATCTGACCAAGAGCTC
285
            |||||||||||||||||||||||||||||||||| ||||||||||||||||||||
Sbjct:     1 AGTTTTCGCCATGTGGACGATGACATGGTCATCCCATACCATCCATCTGACCAAGAGCTC
60
Query:   284 AGTGAGATTAGCTAATAGAATAGTCTAGTCTTTTATAATATATGAAATACATACGCTATC
225
            |||||||||||||||||||||||||||||||||||||||||||||||||||| |||
Sbjct:    61 AGTGAGATTAGCTAATAGAATAGTCTAGTCTTTTATAATATATGAAATACATACACTAAT 120
Query:                                                                224
ATTGTGGGACCCTTTCCTGAAGGGCACCAACAAAGTCATCAGATCGTTTCTCCACCTGGG 165
            ||||||  ||||  || | ||||||||||||||||||||| |||||||   ||||||
Sbjct:   121 ATTGTGAAACCCAATCATAAAGGGCACCAACAAAGTCATCAG-TCGTTTCAGTACCTGGA
179
Query:                                                                164
ACTGCTGGAGTTATAGAGCCCTGCTGTGCCTCTGCCGGAGTAAAGACACTGCTCCTCTCC 105
            ||||| |||||||||| |||||||||||||||| |||||||||||||| |||||||
Sbjct:   180 ACTGCAGGAGTTATAGTGCCCTGCTGTGCCTCTGCAGGAGTAAAGACACTGCACCTCTCC
239
Query:   104 CGCTGCTGAGAATTCACCTCATTTTCAATTGGAGAATGGCTGGTCTGCTGATAGTTTATG
45
            | ||||||| ||||||||||||||||||||||||||||||||||||||||||||| |
Sbjct:   240 CACTGCTGAAAATTCACCTCATTTTCAATTGGAGAATGGCTGGTCTGCTGATAGTTTACG
299
Query:    44 GTCCCTGTCTCAAAATTATTAGAATACAAATTTGAAACCGGATC 1
            |||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   300 GTCCCTGTCTCAAAATTATTAGAATACAAATTTGAAACCGGATC 343
```

MPGR - Retrieved protein sequence

```
>cra_locus_5366_iso_3_len_1889_ver_3
MKVTEKDMSSCFDDGKFFPPGFRFHPTDEELVLYYLKRKICRRRHRLDVIGETDVYKWDPEELP
EISKLKTGDRQWFFFSPRDRKYPNGARSNRATRHGYWKATGKDRIITCNSRPVGVKKTLVFYRG
RAPTGERTDWVMHEYTMDEDELKRCPAKEYYALYKVFKKSGPGPKNGEQYGAPFREEDWADEVE
PNGLVERCKSVEQVAECVPVDDNRINSQLQSQLEDLEEFMNRIADDAILEPPPVDDFAYALGEL
VREEEAQSNIVDQSSKEYNLPEQSIVVLPHCQQYDVQASYDLTQSATSQFQLHETSEVTSAPKG
HIPEPYIVEEDFLEDFLELDDLMGPDPSVQRSHKQAEAHGNEDSIDELDGLFEFELYKDASLVL
SEVGPGGEGQLCQTYVNNMVSKAADPVSNLYSNNFETGTVNYQQTSHSPIENEVNFQQWERCSV
FTPAEAQQGTITPAVPGTETTDDFVGALYDWVSQY
```

MPGR- Retrieved nucleotide sequence

```
>cra_locus_5366_iso_3_len_1889_ver_3
TCAGACCTTCATTTTCTTTCGTTCTTGAGAAATTCGGCTCGATTCTTATTCTTTTATTCATATA
GACCGTGATAGAGTAAAAATTTTGGGAGATACTTTAGTAATTAAAGTTGATTCAATGAAGGTTA
CGGAGAAGGATATGTCTTCTTGTTTTGACGATGGTAAATTTTTCCCGCCGGGATTTAGATTCCA
CCCGACGGATGAAGAGCTTGTTCTGTACTATTTAAAGAGGAAGATCTGCCGGCGGCGACACAGG
CTGGACGTTATCGGCGAGACTGACGTCTACAAGTGGGACCCAGAAGAGTTGCCTGAAATATCTA
```

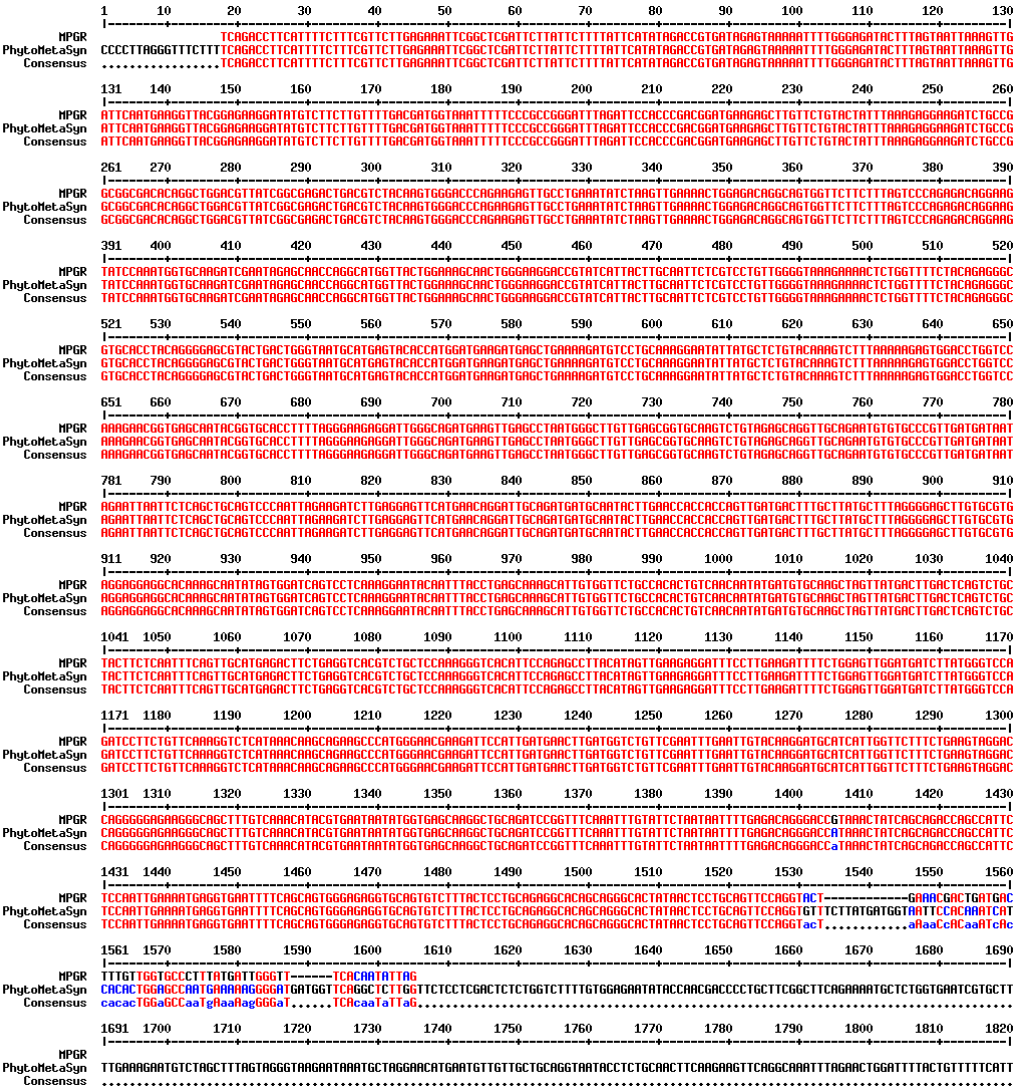
AGTTGAAAACTGGAGACAGGCAGTGGTTCTTCTTTAGTCCCAGAGACAGGAAGTATCCAAATGG
TGCAAGATCGAATAGAGCAACCAGGCATGGTTACTGGAAAGCAACTGGGAAGGACCGTATCATT
ACTTGCAATTCTCGTCCTGTTGGGGTAAAGAAAACTCTGGTTTTCTACAGAGGGCGTGCACCTA
CAGGGGAGCGTACTGACTGGGTAATGCATGAGTACACCATGGATGAAGATGAGCTGAAAAGATG
TCCTGCAAAGGAATATTATGCTCTGTACAAAGTCTTTAAAAAGAGTGGACCTGGTCCAAAGAAC
GGTGAGCAATACGGTGCACCTTTTAGGGAAGAGGATTGGGCAGATGAAGTTGAGCCTAATGGGC
TTGTTGAGCGGTGCAAGTCTGTAGAGCAGGTTGCAGAATGTGTGCCCGTTGATGATAATAGAAT
TAATTCTCAGCTGCAGTCCCAATTAGAAGATCTTGAGGAGTTCATGAACAGGATTGCAGATGAT
GCAATACTTGAACCACCACCAGTTGATGACTTTGCTTATGCTTTAGGGGAGCTTGTGCGTGAGG
AGGAGGCACAAAGCAATATAGTGGATCAGTCCTCAAAGGAATACAATTTACCTGAGCAAAGCAT
TGTGGTTCTGCCACACTGTCAACAATATGATGTGCAAGCTAGTTATGACTTGACTCAGTCTGCT
ACTTCTCAATTTCAGTTGCATGAGACTTCTGAGGTCACGTCTGCTCCAAAGGGTCACATTCCAG
AGCCTTACATAGTTGAAGAGGATTTCCTTGAAGATTTTCTGGAGTTGGATGATCTTATGGGTCC
AGATCCTTCTGTTCAAAGGTCTCATAAACAAGCAGAAGCCCATGGGAACGAAGATTCCATTGAT
GAACTTGATGGTCTGTTCGAATTTGAATTGTACAAGGATGCATCATTGGTTCTTTCTGAAGTAG
GACCAGGGGGAGAAGGGCAGCTTTGTCAAACATACGTGAATAATATGGTGAGCAAGGCTGCAGA
TCCGGTTTCAAATTTGTATTCTAATAATTTTGAGACAGGGACCGTAAACTATCAGCAGACCAGC
CATTCTCCAATTGAAAATGAGGTGAATTTTCAGCAGTGGGAGAGGTGCAGTGTCTTTACTCCTG
CAGAGGCACAGCAGGGCACTATAACTCCTGCAGTTCCAGGTACTGAAACGACTGATGACTTTGT
TGGTGCCCTTTATGATTGGGTTTCACAATATTAG

PhytoMetaSyn - Retrieved nucleotide sequence

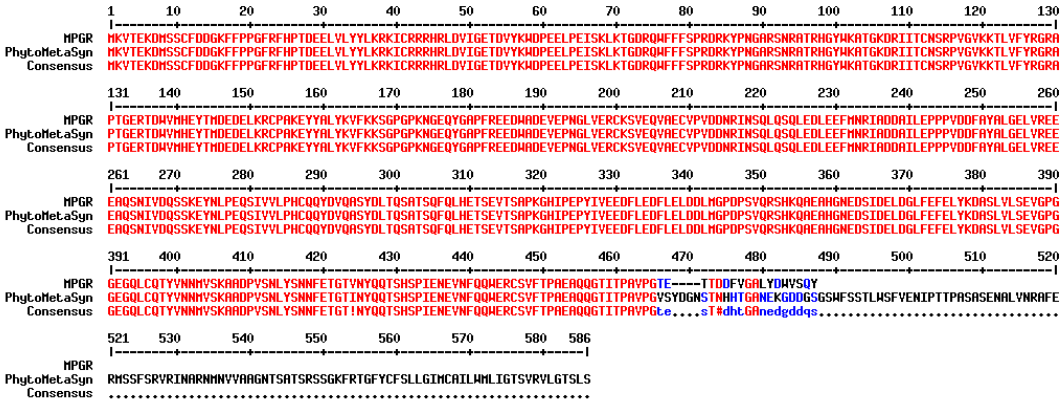
>gnl|MAGPIE|cro.CROWL1VD_mira--CROWL1VD_c4959 No definition line found
CCCCTTAGGGTTTCTTTTCAGACCTTCATTTTCTTTCGTTCTTGAGAAATTCGGCTCGATTCTT
ATTCTTTTATTCATATAGACCGTGATAGAGTAAAAATTTTGGGAGATACTTTAGTAATTAAAGT
TGATTCAATGAAGGTTACGGAGAAGGATATGTCTTCTTGTTTTGACGATGGTAAATTTTTCCCG
CCGGGATTTAGATTCCACCCGACGGATGAAGAGCTTGTTCTGTACTATTTAAAGAGGAAGATCT
GCCGGCGGCGACACAGGCTGGACGTTATCGGCGAGACTGACGTCTACAAGTGGGACCCAGAAGA
GTTGCCTGAAATATCTAAGTTGAAAACTGGAGACAGGCAGTGGTTCTTCTTTAGTCCCAGAGAC
AGGAAGTATCCAAATGGTGCAAGATCGAATAGAGCAACCAGGCATGGTTACTGGAAAGCAACTG
GGAAGGACCGTATCATTACTTGCAATTCTCGTCCTGTTGGGGTAAAGAAAACTCTGGTTTTCTA
CAGAGGGCGTGCACCTACAGGGGAGCGTACTGACTGGGTAATGCATGAGTACACCATGGATGAA
GATGAGCTGAAAAGATGTCCTGCAAAGGAATATTATGCTCTGTACAAAGTCTTTAAAAAGAGTG
GACCTGGTCCAAAGAACGGTGAGCAATACGGTGCACCTTTTAGGGAAGAGGATTGGGCAGATGA
AGTTGAGCCTAATGGGCTTGTTGAGCGGTGCAAGTCTGTAGAGCAGGTTGCAGAATGTGTGCCC
GTTGATGATAATAGAATTAATTCTCAGCTGCAGTCCCAATTAGAAGATCTTGAGGAGTTCATGA
ACAGGATTGCAGATGATGCAATACTTGAACCACCACCAGTTGATGACTTTGCTTATGCTTTAGG
GGAGCTTGTGCGTGAGGAGGAGGCACAAAGCAATATAGTGGATCAGTCCTCAAAGGAATACAAT
TTACCTGAGCAAAGCATTGTGGTTCTGCCACACTGTCAACAATATGATGTGCAAGCTAGTTATG
ACTTGACTCAGTCTGCTACTTCTCAATTTCAGTTGCATGAGACTTCTGAGGTCACGTCTGCTCC
AAAGGGTCACATTCCAGAGCCTTACATAGTTGAAGAGGATTTCCTTGAAGATTTTCTGGAGTTG
GATGATCTTATGGGTCCAGATCCTTCTGTTCAAAGGTCTCATAAACAAGCAGAAGCCCATGGGA
ACGAAGATTCCATTGATGAACTTGATGGTCTGTTCGAATTTGAATTGTACAAGGATGCATCATT
GGTTCTTTCTGAAGTAGGACCAGGGGGAGAAGGGCAGCTTTGTCAAACATACGTGAATAATATG
GTGAGCAAGGCTGCAGATCCGGTTTCAAATTTGTATTCTAATAATTTTGAGACAGGGACCATAA
ACTATCAGCAGACCAGCCATTCTCCAATTGAAAATGAGGTGAATTTTCAGCAGTGGGAGAGGTG

CAGTGTCTTTACTCCTGCAGAGGCACAGCAGGGCACTATAACTCCTGCAGTTCCAGGTGTTTCT
TATGATGGTAATTCCACAAATCATCACACTGGAGCCAATGAAAAAGGGGATGATGGTTCAGGCT
CTTGGTTCTCCTCGACTCTCTGGTCTTTTGTGGAGAATATACCAACGACCCCTGCTTCGGCTTC
AGAAAATGCTCTGGTGAATCGTGCTTTTGAAAGAATGTCTAGCTTTAGTAGGGTAAGAATAAAT
GCTAGGAACATGAATGTTGTTGCTGCAGGTAATACCTCTGCAACTTCAAGAAGTTCAGGCAAAT
TTAGAACTGGATTTTACTGTTTTTCATTATTGGGTATCATGTGTGCAATCTTGTGGATGTTGAT
AGGAACATCTGTGAGAGTCCTTGGAACATCTTTATCCTAATAAATATGTATATCCTATATCCTA
GTTTTAACTGTTTTGTCAAAATAGTTTGAAGGCCAAAGTTGCTGCTGAAGAATGAGGCTTTCTA
TGCGCTTGATTATGCCTTAAAAAATGTAATACATATTCATGCAACAGGCCTTGTTTTAATGAGG
TGTATCTAGATCTGGTTCCGATTTGGATTGGAAGATAATTTGAAATCATATGACTTCTTATGAC
AATGTAAATATTTGCTGATCTGCAGCTAAAGAATCTTATAAACATTCTTCTGCACCTTTTTCT

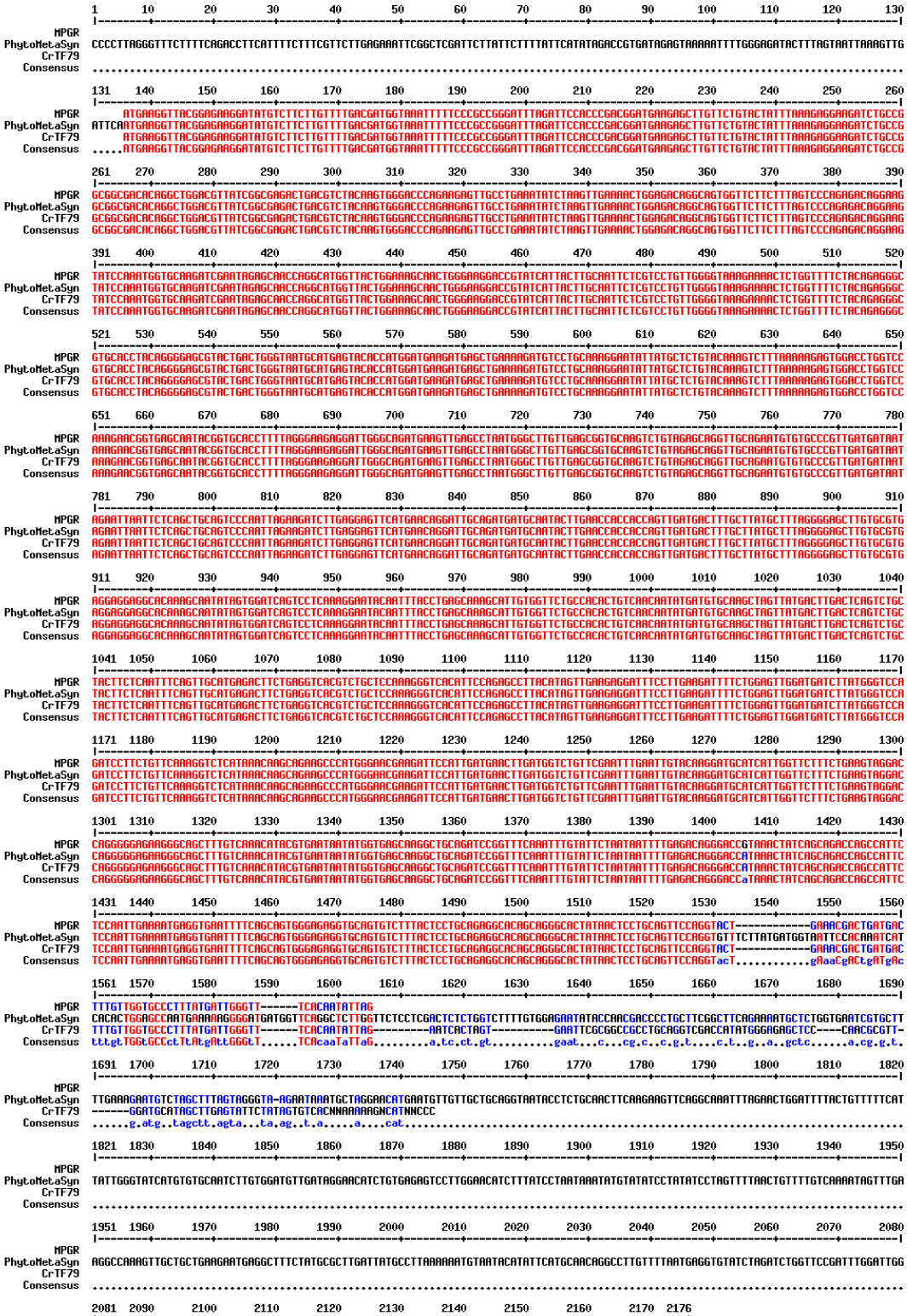
Alignment of the retrieved nucleotide sequences retrieved from MPGR and PhytoMetaSyn databases



Alignment between the proteins translated from the nucleotide sequences retrieved from MPGR and PhytoMetaSyn databases



Alignment between the nucleotide sequences retrieved from MPGR and PhytoMetaSyn databases, and from sequenced *CrTF79*.



CrTF152

Tag and MPGR best hit alignment

```
>cra_locus_4787_iso_1_len_2989_ver_3
      Length = 2989

 Minus Strand HSPs:

 Score = 1967 (301.2 bits), Expect = 2.1e-84, P = 2.1e-84
 Identities = 395/397 (99%), Positives = 395/397 (99%), Strand = Minus / Plus

Query:                                                                397
TTAACCAATTTACAGAGGTGCCAAACCTCCTCGAGAGCGCGATCAAGCGTGCCGAGAGGC 338
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:                                                                219
TTAACCAATTTACAGAGGTGCCAAACCTCCTCGAGAGCGCGATCAAGCGTGCCGAGAGGC 278

Query:     337  TACCTAACATACTTCTTTCTCCCTGTCAACTTTTTACATAACTGTAACAATGACAAAAAA
278
            |||||||  ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   279 TACCTAATTTACTTCTTTCTCCCTGTCAACTTTTTACATAACTGTAACAATGACAAAAAA 338

Query:     277  AATTCATAATAATAATAATAATAATAGTAATTAGAGAACACCTTTCGAGGTCTTCTTCTA
218
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   339 AATTCATAATAATAATAATAATAATAGTAATTAGAGAACACCTTTCGAGGTCTTCTTCTA 398

Query:    217  TTTTACTCTCTGTTTGTGCTCTCATATCCTTCCCTTTTCAAAGAAGGCGGCAAGCCTCTG
158
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:    399  TTTTACTCTCTGTTTGTGCTCTCATATCCTTCCCTTTTCAAAGAAGGCGGCAAGCCTCTG
458

Query:                                                                157
TCGGCTGTTTGAACTGCTGCTTTACTTGCTGCTGCGACCAGTAAGCATGTGCAGTCAGGA 98
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:    459 TCGGCTGTTTGAACTGCTGCTTTACTTGCTGCTGCGACCAGTAAGCATGTGCAGTCAGGA
518

Query:                                                                 97
CTCGGTCAAGGAGCTCTTGTGGCACAGGCTCACCTCTCCTCTGTTGAGGGGATATTCTTG 38
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:    519 CTCGGTCAAGGAGCTCTTGTGGCACAGGCTCACCTCTCCTCTGTTGAGGGGATATTCTTG
578

Query:    37 CAGTGTGCACCAAGGTTTTCCACTTGTCCTTGAGATC 1
            |||||||||||||||||||||||||||||||||||||
Sbjct:   579 CAGTGTGCACCAAGGTTTTCCACTTGTCCTTGAGATC 615
```

MPGR - Retrieved protein sequence

```
>cra_locus_4787_iso_1_len_2989_ver_3
MVLQNRLDYGFNGYQVPPTPRAARSARKRGIYRRKVDDSQMCAFDLLATIAGKLLTEGGTSPSS
VDMGSGKEKSEAVKDSYKEDQHEEDKHPNAKFCKEDTCGRSFIVSELVLQAPILNNGSKELPPT
QNDVCSGPASGITSSESSEKVGSDEHLANDEIKAGTGHCSGKADIELFGCRLSPGCKLEDESKK
QIKVEPNDLKLSISTRDDVCNSDCAVAWDQKPYTLGSSDDSVKFSQTRDSISCGSLPFNRENVK
LVTRDDDEKSAGCLQSSTRNKAFRLSPHTGDRKIRKFLVSKYWKVPNLKDDEHLKTDADAKHAY
QNAKNGNKRQRSQRDYPFKKRKLYGCTSVSNSDGGISSDETCKSPRKGSAGDSSDSMLPRVTAI
SSSAAGEHSSLHSKNSPVKLKIKSFRVPEFFIEIPETATVGSLKKTVMEAVTAILSGGLHVGVL
```

LQGKKIRDDNKTLLQSGISHDNKPDALGFTLEPNTSQSPHCLNPEDHPFLLPCDAPEPLSRYPP
VSVTNHNAVHRGTSDASADPPGTNLSNFIESDHDSAPSPPDMSLEKSTVDSRALVAVPAMNVGP
LTVVPMRKSKRSEAAQRRIRRPFSVSEVEALVQAVEKLGTGRWRDVKLRAFDNAKHRTYVDLKD
KWKTLVHTARISPQQRRGEPVPQELLDRVLTAHAYWSQQQVKQQFKQPTEACRLL

MPGR- Retrieved nucleotide sequence

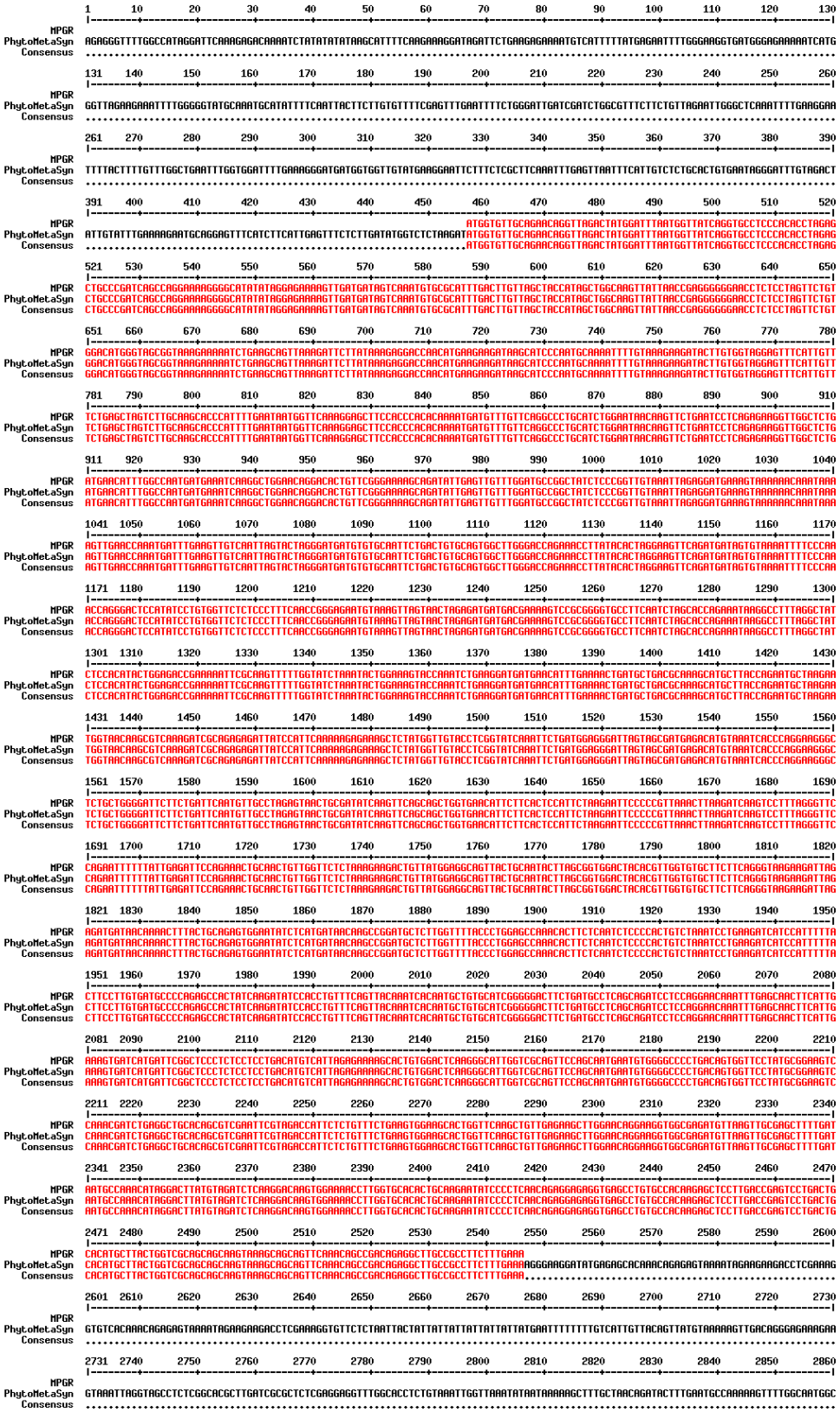
>cra_locus_4787_iso_1_len_2989_ver_3
TTTCTCTTGATATGGTCTCTAAGATATGGTGTTGCAGAACAGGTTAGACTATGGATTTAATGGT
TATCAGGTGCCTCCCACACCTAGAGCTGCCCGATCAGCCAGGAAAAGGGGCATATATAGGAGAA
AAGTTGATGATAGTCAAATGTGCGCATTTGACTTGTTAGCTACCATAGCTGGCAAGTTATTAAC
CGAGGGGGGAACCTCTCCTAGTTCTGTGGACATGGGTAGCGGTAAAGAAAAATCTGAAGCAGTT
AAAGATTCTTATAAAGAGGACCAACATGAAGAAGATAAGCATCCCAATGCAAAATTTTGTAAAG
AAGATACTTGTGGTAGGAGTTTCATTGTTTCTGAGCTAGTCTTGCAAGCACCCATTTTGAATAA
TGGTTCAAAGGAGCTTCCACCCACACAAAATGATGTTTGTTCAGGCCCTGCATCTGGAATAACA
AGTTCTGAATCCTCAGAGAAGGTTGGCTCTGATGAACATTTGGCCAATGATGAAATCAAGGCTG
GAACAGGACACTGTTCGGGAAAAGCAGATATTGAGTTGTTTGGATGCCGGCTATCTCCCGGTTG
TAAATTAGAGGATGAAAGTAAAAAACAAATAAAAGTTGAACCAAATGATTTGAAGTTGTCAATT
AGTACTAGGGATGATGTGTGCAATTCTGACTGTGCAGTGGCTTGGGACCAGAAACCTTATACAC
TAGGAAGTTCAGATGATAGTGTAAAATTTTCCCAAACCAGGGACTCCATATCCTGTGGTTCTCT
CCCTTTCAACCGGGAGAATGTAAAGTTAGTAACTAGAGATGATGACGAAAAGTCCGCGGGGTGC
CTTCAATCTAGCACCAGAAATAAGGCCTTTAGGCTATCTCCACATACTGGAGACCGAAAAATTC
GCAAGTTTTTGGTATCTAAATACTGGAAAGTACCAAATCTGAAGGATGATGAACATTTGAAAAC
TGATGCTGACGCAAAGCATGCTTACCAGAATGCTAAGAATGGTAACAAGCGTCAAAGATCGCAG
AGAGATTATCCATTCAAAAAGAGAAAGCTCTATGGTTGTACCTCGGTATCAAATTCTGATGGAG
GGATTAGTAGCGATGAGACATGTAAATCACCCAGGAAGGGCTCTGCTGGGGATTCTTCTGATTC
AATGTTGCCTAGAGTAACTGCGATATCAAGTTCAGCAGCTGGTGAACATTCTTCACTCCATTCT
AAGAATTCCCCCGTTAAACTTAAGATCAAGTCCTTTAGGGTTCCAGAATTTTTTATTGAGATTC
CAGAAACTGCAACTGTTGGTTCTCTAAAGAAGACTGTTATGGAGGCAGTTACTGCAATACTTAG
CGGTGGACTACACGTTGGTGTGCTTCTTCAGGGTAAGAAGATTAGAGATGATAACAAAACTTTA
CTGCAGAGTGGAATATCTCATGATAACAAGCCGGATGCTCTTGGTTTTACCCTGGAGCCAAACA
CTTCTCAATCTCCCCACTGTCTAAATCCTGAAGATCATCCATTTTTACTTCCTTGTGATGCCCC
AGAGCCACTATCAAGATATCCACCTGTTTCAGTTACAAATCACAATGCTGTGCATCGGGGGACT
TCTGATGCCTCAGCAGATCCTCCAGGAACAAATTTGAGCAACTTCATTGAAAGTGATCATGATT
CGGCTCCCTCTCCTCCTGACATGTCATTAGAGAAAAGCACTGTGGACTCAAGGGCATTGGTCGC
AGTTCCAGCAATGAATGTGGGGCCCCTGACAGTGGTTCCTATGCGGAAGTCCAAACGATCTGAG
GCTGCACAGCGTCGAATTCGTAGACCATTCTCTGTTTCTGAAGTGGAAGCACTGGTTCAAGCTG
TTGAGAAGCTTGGAACAGGAAGGTGGCGAGATGTTAAGTTGCGAGCTTTTGATAATGCCAAACA
TAGGACTTATGTAGATCTCAAGGACAAGTGGAAAACCTTGGTGCACACTGCAAGAATATCCCCT
CAACAGAGGAGAGGTGAGCCTGTGCCACAAGAGCTCCTTGACCGAGTCCTGACTGCACATGCTT
ACTGGTCGCAGCAGCAAGTAAAGCAGCAGTTCAAACAGCCGACAGAGGCTTGCCGCCTTCTTTG
AAA

PhytoMetaSyn - Retrieved nucleotide sequence

>gnl|MAGPIE|cro.CROWL1VD_velvet--Singlet5629 No definition line found
AGAGGGTTTTGGCCATAGGATTCAAAGAGACAAAATCTATATATATAAGCATTTTCAAGAAAGG
ATAGATTCTGAAGAGAAAATGTCATTTTTATGAGAATTTTGGGAAGGTGATGGGAGAAAAATCA

TGGGTTAGAAGAAATTTTGGGGGTATGCAAATGCATATTTTCAATTACTTCTTGTGTTTTCGAG
TTTGAATTTTCTGGGATTGATCGATCTGGCGTTTCTTCTGTTAGAATTGGGCTCAAATTTTGAA
GGAATTTTACTTTTGTTTGGCTGAATTTGGTGGATTTTGAAAGGGATGATGGTGGTTGTATGAA
GGAATTCTTTCTCGCTTCAAATTTGAGTTAATTTCATTGTCTCTGCACTGTGAATAGGGATTTG
TAGACTATTGTATTTGAAAAGAATGCAGGAGTTTCATCTTCATTGAGTTTCTCTTGATATGGTC
TCTAAGATATGGTGTTGCAGAACAGGTTAGACTATGGATTTAATGGTTATCAGGTGCCTCCCAC
ACCTAGAGCTGCCCGATCAGCCAGGAAAAGGGGCATATATAGGAGAAAAGTTGATGATAGTCAA
ATGTGCGCATTTGACTTGTTAGCTACCATAGCTGGCAAGTTATTAACCGAGGGGGGAACCTCTC
CTAGTTCTGTGGACATGGGTAGCGGTAAAGAAAAATCTGAAGCAGTTAAAGATTCTTATAAAGA
GGACCAACATGAAGAAGATAAGCATCCCAATGCAAAATTTTGTAAAGAAGATACTTGTGGTAGG
AGTTTCATTGTTTCTGAGCTAGTCTTGCAAGCACCCATTTTGAATAATGGTTCAAAGGAGCTTC
CACCCACACAAAATGATGTTTGTTCAGGCCCTGCATCTGGAATAACAAGTTCTGAATCCTCAGA
GAAGGTTGGCTCTGATGAACATTTGGCCAATGATGAAATCAAGGCTGGAACAGGACACTGTTCG
GGAAAAGCAGATATTGAGTTGTTTGGATGCCGGCTATCTCCCGGTTGTAAATTAGAGGATGAAA
GTAAAAAACAAATAAAAGTTGAACCAAATGATTTGAAGTTGTCAATTAGTACTAGGGATGATGT
GTGCAATTCTGACTGTGCAGTGGCTTGGGACCAGAAACCTTATACACTAGGAAGTTCAGATGAT
AGTGTAAAATTTTCCCAAACCAGGGACTCCATATCCTGTGGTTCTCTCCCTTTCAACCGGGAGA
ATGTAAAGTTAGTAACTAGAGATGATGACGAAAAGTCCGCGGGGTGCCTTCAATCTAGCACCAG
AAATAAGGCCTTTAGGCTATCTCCACATACTGGAGACCGAAAAATTCGCAAGTTTTTGGTATCT
AAATACTGGAAAGTACCAAATCTGAAGGATGATGAACATTTGAAAACTGATGCTGACGCAAAGC
ATGCTTACCAGAATGCTAAGAATGGTAACAAGCGTCAAAGATCGCAGAGAGATTATCCATTCAA
AAAGAGAAAGCTCTATGGTTGTACCTCGGTATCAAATTCTGATGGAGGGATTAGTAGCGATGAG
ACATGTAAATCACCCAGGAAGGGCTCTGCTGGGGATTCTTCTGATTCAATGTTGCCTAGAGTAA
CTGCGATATCAAGTTCAGCAGCTGGTGAACATTCTTCACTCCATTCTAAGAATTCCCCCGTTAA
ACTTAAGATCAAGTCCTTTAGGGTTCCAGAATTTTTTATTGAGATTCCAGAAACTGCAACTGTT
GGTTCTCTAAAGAAGACTGTTATGGAGGCAGTTACTGCAATACTTAGCGGTGGACTACACGTTG
GTGTGCTTCTTCAGGGTAAGAAGATTAGAGATGATAACAAAACTTTACTGCAGAGTGGAATATC
TCATGATAACAAGCCGGATGCTCTTGGTTTTACCCTGGAGCCAAACACTTCTCAATCTCCCCAC
TGTCTAAATCCTGAAGATCATCCATTTTTACTTCCTTGTGATGCCCCAGAGCCACTATCAAGAT
ATCCACCTGTTTCAGTTACAAATCACAATGCTGTGCATCGGGGGACTTCTGATGCCTCAGCAGA
TCCTCCAGGAACAAATTTGAGCAACTTCATTGAAAGTGATCATGATTCGGCTCCCTCTCCTCCT
GACATGTCATTAGAGAAAAGCACTGTGGACTCAAGGGCATTGGTCGCAGTTCCAGCAATGAATG
TGGGGCCCCTGACAGTGGTTCCTATGCGGAAGTCCAAACGATCTGAGGCTGCACAGCGTCGAAT
TCGTAGACCATTCTCTGTTTCTGAAGTGGAAGCACTGGTTCAAGCTGTTGAGAAGCTTGGAACA
GGAAGGTGGCGAGATGTTAAGTTGCGAGCTTTTGATAATGCCAAACATAGGACTTATGTAGATC
TCAAGGACAAGTGGAAAACCTTGGTGCACACTGCAAGAATATCCCCTCAACAGAGGAGAGGTGA
GCCTGTGCCACAAGAGCTCCTTGACCGAGTCCTGACTGCACATGCTTACTGGTCGCAGCAGCAA
GTAAAGCAGCAGTTCAAACAGCCGACAGAGGCTTGCCGCCTTCTTTGAAAAGGGAAGGATATGA
GAGCACAAACAGAGAGTAAAATAGAAGAAGACCTCGAAAGGTGTCACAAACAGAGAGTAAAATA
GAAGAAGACCTCGAAAGGTGTTCTCTAATTACTATTATTATTATTATTATTATGAATTTTTTTT
GTCATTGTTACAGTTATGTAAAAAGTTGACAGGGAGAAAGAAGTAAATTAGGTAGCCTCTCGGC
ACGCTTGATCGCGCTCTCGAGGAGGTTTGGCACCTCTGTAAATTGGTTAAATATAATAAAAAGC
TTTGCTAACAGATACTTTGAATGCCAAAAAGTTTTGGCAATGGCCGTATGTCCAGTTAGGTTTT
TTTTTTTTACATGAATTCTTTTGATTTTCTTGGGGTTGGAATTTTTTTTTTTTACATGAATTCT
TTTGATTTTCTTGGGGTTGGAAATGGAAAATTGCTGCTCCCTTTTGGTTTTGGAGTGCATTTGG
TTGTAAATTTGTTTAGCAGAAATGTTGAGATTCATTCACTCTTTGTAAGAATGAGATATTTTAT
TAGACAGTTATAAATCTTGTAGTCCCCTCTTTCCATTTTGCTCTTT

Alignment of the retrieved nucleotide sequences retrieved from MPGR and PhytoMetaSyn databases



CrTF194

Tag and MPGR best hit alignment

```
>cra_locus_4594_iso_2_len_2637_ver_3

        Length = 2637

  Minus Strand HSPs:

 Score = 2778 (422.9 bits), Expect = 5.5e-121, P = 5.5e-121

 Identities = 556/559 (99%), Positives = 558/559 (99%), Strand = Minus / Plus

Query:   559 YSATTAACAAACCCCTGTAAATTGTTTGGTTGTGTCTGGGGACTATAAAGTTCCACGGGA 500

             ++||||||||||||||||||||||||||||||||||||||||||||||||||||||||||

Sbjct:  1308 CCATTAACAAACCCCTGTAAATTGTTTGGTTGTGTCTGGGGACTATAAAGTTCCACGGGA 1367

Query:   499 TTGCCTGGCTTGACATGATTGAACTGTGGCCACGCGGTGTTATTCAGACCATTGCACGTG 440

             |||||||||||||||||||||||||||||||||||||||||||||||| |||||||||

Sbjct:  1368 TTGCCTGGCTTGACATGATTGAACTGTGGCCACGCGGTGTTATTCAGACCGTTGCACGTG 1427

Query:   439 TTTGGAAACGTAGTAATTCTGTTTCCATTTGGATAAGCATCCCACTTTCTTTCGTCTGGC 380

             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||

Sbjct:  1428 TTTGGAAACGTAGTAATTCTGTTTCCATTTGGATAAGCATCCCACTTTCTTTCGTCTGGC 1487

Query:   379 TTTCTGACGACAAGCTTTTCCCGACTTTGAGCATGAATTGAATTCAAATCTTGACCAAAA 320

             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||

Sbjct:  1488 TTTCTGACGACAAGCTTTTCCCGACTTTGAGCATGAATTGAATTCAAATCTTGACCAAAA 1547

Query:   319 ATCTTGTGAACGCTGTCTGGCCGTTGCCCTATCCCCAAATTCATGTAAGCTGCATTGCCA 260

             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||

Sbjct:  1548 ATCTTGTGAACGCTGTCTGGCCGTTGCCCTATCCCCAAATTCATGTAAGCTGCATTGCCA 1607

Query:   259 TCCACGTCGGCTGCCACTGTCACAGGTGCAACTTGTTTGGCCCTAATTAGTGACGAATAC 200

             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||

Sbjct:  1608 TCCACGTCGGCTGCCACTGTCACAGGTGCAACTTGTTTGGCCCTAATTAGTGACGAATAC 1667

Query:   199 AATGAGAAGGAAGACCTAACGGACTGCATGACCTCAAAACTTTCCGGAATGCATCTAACT 140

             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||

Sbjct:  1668 AATGAGAAGGAAGACCTAACGGACTGCATGACCTCAAAACTTTCCGGAATGCATCTAACT 1727

Query:   139 GAACCTAATTCAACTACCCCAACATCAGTAGGAATCAGCACAATAGTTCGCATACCAGCC 80

             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||

Sbjct:  1728 GAACCTAATTCAACTACCCCAACATCAGTAGGAATCAGCACAATAGTTCGCATACCAGCC 1787

Query:    79 GACTTCGCTAGAAATGACCTAACACAATAATCTGTAGATGACTTCAATGAATCAGATAAC 20

             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
```

```
Sbjct:  1788 GACTTCGCTAGAAATGACCTAACACAATAATCTGTAGATGACTTCAATGAATCAGATAAC 1847

Query:    19 CAAAAAGGCTTCCCAGATC 1

             ||||||||||||||||||

Sbjct:  1848 CAAAAAGGCTTCCCAGATC 1866
```

MPGR - Retrieved protein sequence

```
>cra_locus_4594_iso_2_len_2637_ver_3
MVAAVLGTKAFDFLFSNTVSADGSLMVMGNDENLQNKLSDLVEHPNAANFSWNYAIFWQISRSK
SGELVLGWGDGCCREPRDGEESEFTQILSLRLEDETQQRMRKRVLQKLHTCFGGTEEDSYAFGL
DRVTDTEMFFLASMYFSFPRGEGGPGKCFGSGKPFWLSDSLKSSTDYCVRSFLAKSAGMRTIVL
IPTDVGVVELGSVRCIPESFEVMQSVRSSFSLYSSLIRAKQVAPVTVAADVDGNAAYMNLGIGQ
RPDSVHKIFGQDLNSIHAQSREKLVVRKPDERKWDAYPNGNRITTFPNTCNGLNNTAWPQFNHV
KPGNPVELYSPQTQPNNLQGFVNGASRVNNFQHQKPAQMQIDFTGATSRPVVSRPVSAESEHSD
IEASCREESAVLAADEKRPRKRGRKPANGREEPLNHVEAERQRREKLNQRFYALRAVVPNISKM
DKASLLGDAIAYITELQKKLRDLESEKEKVGSTSRDALAFEKNASSGVENQVPEIDIEAARDEV
IVKVTCPLDAHPVSGVIQAFKETQVNVLDSKLATGNETVFHTFVVKSQGSDKLTKEKLIAAMSR
DREATSQSS
```

MPGR- Retrieved nucleotide sequence

```
>cra_locus_4594_iso_2_len_2637_ver_3
TATCAAAATTCTACAGTACAGGCAAAGAGTTCTTCTTCTTCTTCTTCTGGAGGTGGATTCTTCT
GGAAATGAAAATAGGAGTCGGTATGGGAGGTGTTGTTTGGACTGATGAAGATAAAACAATGGTG
GCGGCGGTTTTGGGGACTAAGGCTTTTGATTTCTTGTTTTCCAATACTGTTTCAGCTGACGGTT
CATTAATGGTTATGGGGAATGATGAAAATTTGCAGAATAAGCTTTCAGATCTCGTTGAACATCC
AAACGCTGCCAATTTTAGTTGGAATTATGCAATTTTCTGGCAAATTTCGCGGTCCAAGTCCGGC
GAATTGGTATTAGGATGGGGTGATGGCTGTTGCAGAGAACCTAGAGATGGGGAGGAATCCGAAT
TTACCCAAATTTTAAGTCTCCGGCTTGAGGATGAAACTCAACAGAGGATGAGGAAAAGAGTTCT
TCAGAAACTGCATACATGCTTTGGTGGAACAGAAGAGGATAGTTATGCTTTTGGATTGGATAGG
GTTACTGATACTGAGATGTTCTTCCTGGCATCCATGTACTTTTCATTTCCTCGAGGAGAAGGAG
GTCCGGGGAAGTGTTTTGGATCTGGGAAGCCTTTTTGGTTATCTGATTCATTGAAGTCATCTAC
AGATTATTGTGTTAGGTCATTTCTAGCGAAGTCGGCTGGTATGCGAACTATTGTGCTGATTCCT
ACTGATGTTGGGGTAGTTGAATTAGGTTCAGTTAGATGCATTCCGGAAAGTTTTGAGGTCATGC
AGTCCGTTAGGTCTTCCTTCTCATTGTATTCGTCACTAATTAGGGCCAAACAAGTTGCACCTGT
GACAGTGGCAGCCGACGTGGATGGCAATGCAGCTTACATGAATTTGGGGATAGGGCAACGGCCA
GACAGCGTTCACAAGATTTTTGGTCAAGATTTGAATTCAATTCATGCTCAAAGTCGGGAAAAGC
TTGTCGTCAGAAAGCCAGACGAAAGAAAGTGGGATGCTTATCCAAATGGAAACAGAATTACTAC
GTTTCCAAACACGTGCAACGGTCTGAATAACACCGCGTGGCCACAGTTCAATCATGTCAAGCCA
GGCAATCCCGTGGAACTTTATAGTCCCCAGACACAACCAAACAATTTACAGGGGTTTGTTAATG
GAGCAAGTCGTGTAAACAATTTTCAGCATCAAAAGCCAGCTCAAATGCAAATTGACTTTACAGG
AGCGACATCAAGACCTGTAGTTTCAAGACCTGTAAGTGCTGAGTCTGAGCATTCAGATATTGAG
GCTTCATGCAGAGAAGAGAGTGCAGTTCTAGCTGCTGATGAAAAGAGGCCTCGGAAGCGTGGTC
GGAAACCTGCCAATGGAAGAGAAGAGCCCCTCAATCATGTCGAGGCAGAGAGGCAGCGAAGAGA
GAAGCTGAATCAGCGGTTTTACGCACTGAGAGCTGTTGTACCCAATATTTCTAAAATGGATAAA
GCATCCCTTTTGGGAGATGCCATTGCTTACATAACCGAGCTGCAGAAGAAGCTTAGGGACCTGG
```

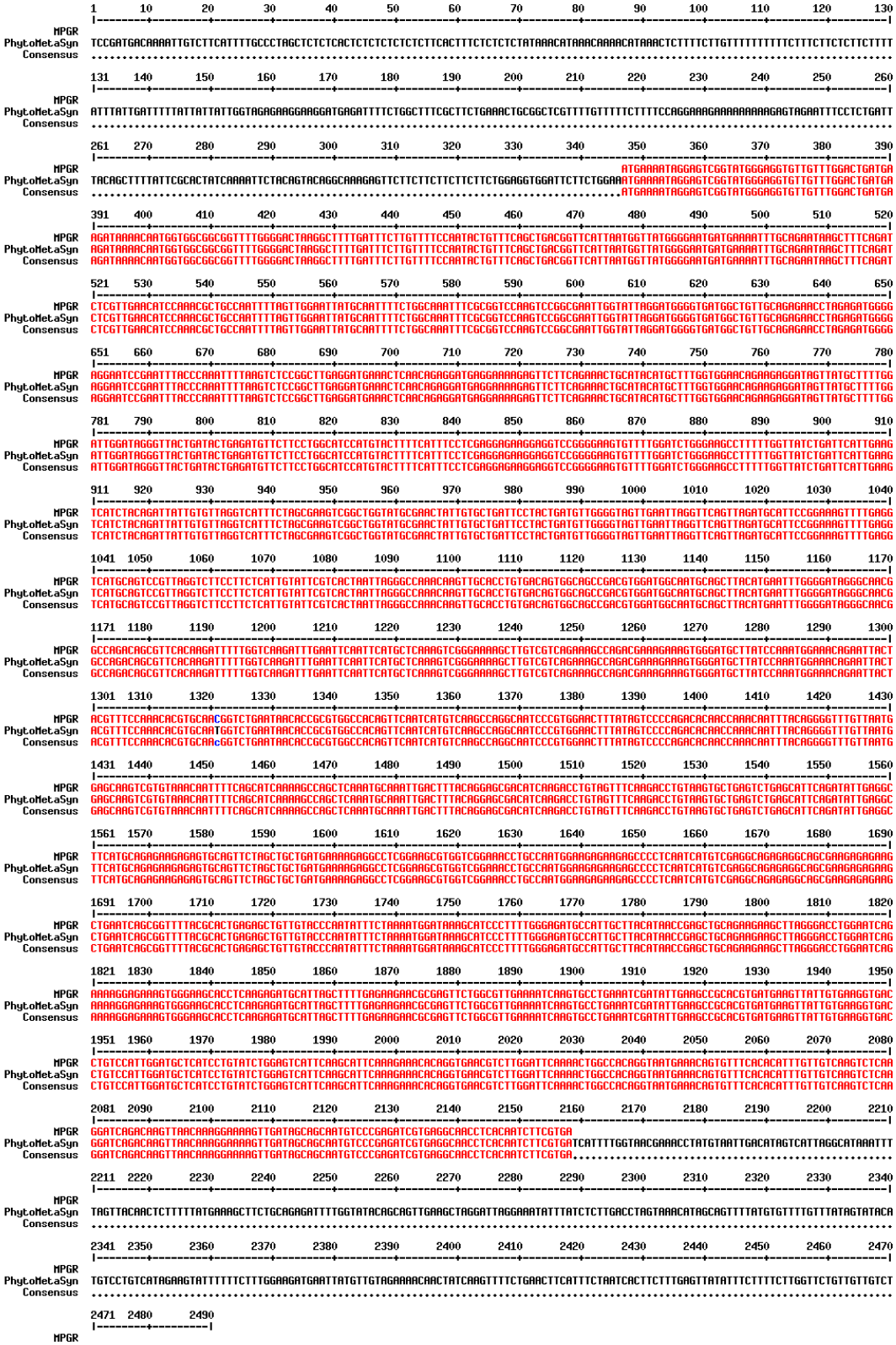
AATCAGAAAAGGAGAAAGTGGGAAGCACCTCAAGAGATGCATTAGCTTTTGAGAAGAACGCGAG
TTCTGGCGTTGAAAATCAAGTGCCTGAAATCGATATTGAAGCCGCACGTGATGAAGTTATTGTG
AAGGTGACCTGTCCATTGGATGCTCATCCTGTATCTGGAGTCATTCAAGCATTCAAAGAAACAC
AGGTGAACGTCTTGGATTCAAAACTGGCCACAGGTAATGAAACAGTGTTTCACACATTTGTTGT
CAAGTCTCAAGGATCAGACAAGTTAACAAAGGAAAAGTTGATAGCAGCAATGTCCCGAGATCGT
GAGGCAACCTCACAATCTTCGTGA

PhytoMetaSyn - Retrieved nucleotide sequence

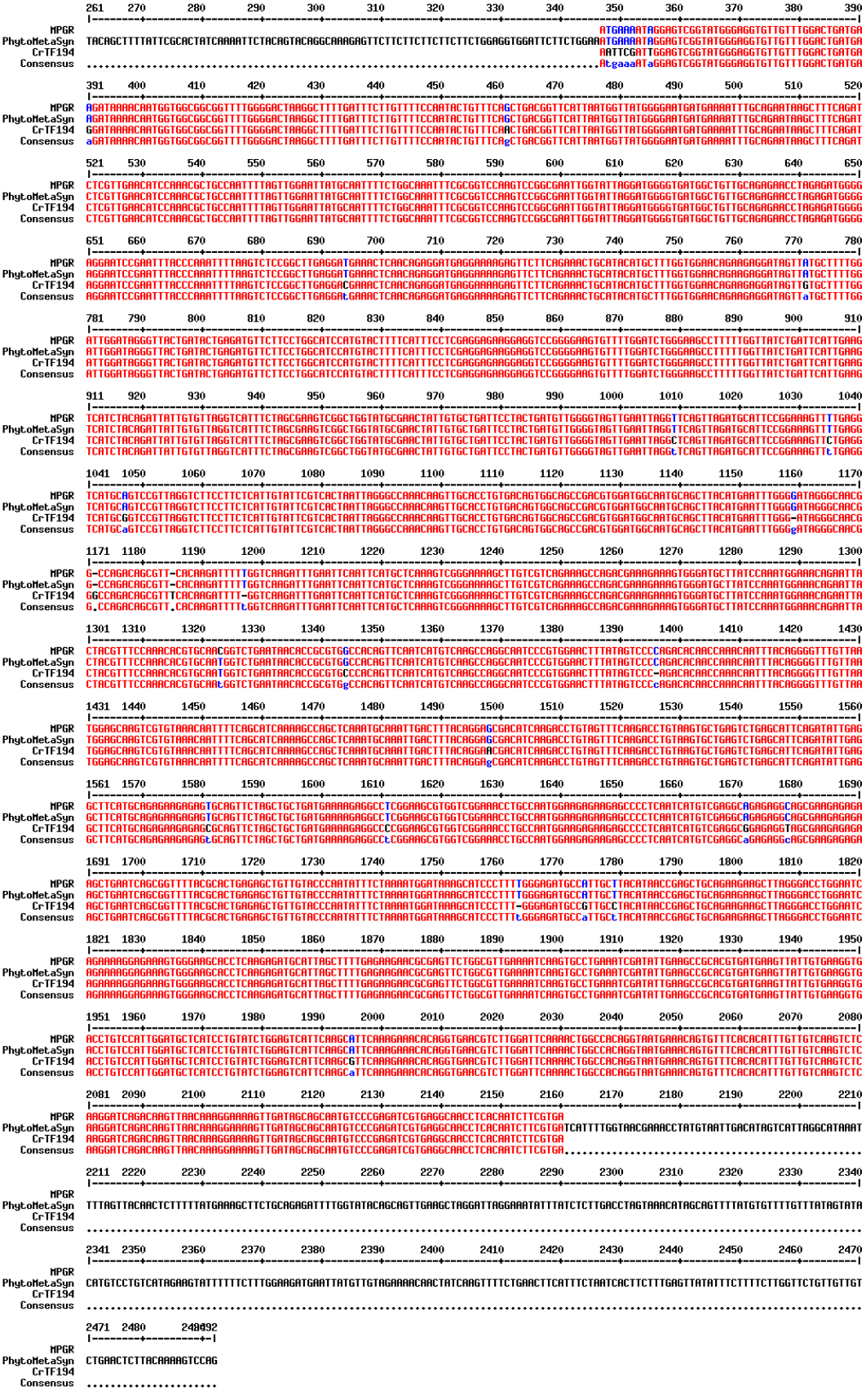
>gnl|MAGPIE|cro.CROLV1VD_velvet--Contig14815 No definition line found
TCCGATGACAAAATTGTCTTCATTTTGCCCTAGCTCTCTCACTCTCTCTCTCTCTTCACTTTCT
CTCTCTATAAACATAAACAAAACATAAACTCTTTTCTTGTTTTTTTTTTCTTTCTTCTCTTCTT
TTATTTATTGATTTTTATTATTATTGGTAGAGAAGGAAGGATGAGATTTTCTGGCTTTCGCTTC
TGAAACTGCGGCTCGTTTTGTTTTTCTTTTCCAGGAAAGAAAAAAAAAGAGTAGAATTTCCTCT
GATTTACAGCTTTTATTCGCACTATCAAAATTCTACAGTACAGGCAAAGAGTTCTTCTTCTTCT
TCTTCTGGAGGTGGATTCTTCTGGAAATGAAAATAGGAGTCGGTATGGGAGGTGTTGTTTGGAC
TGATGAAGATAAAACAATGGTGGCGGCGGTTTTGGGGACTAAGGCTTTTGATTTCTTGTTTTCC
AATACTGTTTCAGCTGACGGTTCATTAATGGTTATGGGGAATGATGAAAATTTGCAGAATAAGC
TTTCAGATCTCGTTGAACATCCAAACGCTGCCAATTTTAGTTGGAATTATGCAATTTTCTGGCA
AATTTCGCGGTCCAAGTCCGGCGAATTGGTATTAGGATGGGGTGATGGCTGTTGCAGAGAACCT
AGAGATGGGGAGGAATCCGAATTTACCCAAATTTTAAGTCTCCGGCTTGAGGATGAAACTCAAC
AGAGGATGAGGAAAAGAGTTCTTCAGAAACTGCATACATGCTTTGGTGGAACAGAAGAGGATAG
TTATGCTTTTGGATTGGATAGGGTTACTGATACTGAGATGTTCTTCCTGGCATCCATGTACTTT
TCATTTCCTCGAGGAGAAGGAGGTCCGGGGAAGTGTTTTGGATCTGGGAAGCCTTTTTGGTTAT
CTGATTCATTGAAGTCATCTACAGATTATTGTGTTAGGTCATTTCTAGCGAAGTCGGCTGGTAT
GCGAACTATTGTGCTGATTCCTACTGATGTTGGGGTAGTTGAATTAGGTTCAGTTAGATGCATT
CCGGAAAGTTTTGAGGTCATGCAGTCCGTTAGGTCTTCCTTCTCATTGTATTCGTCACTAATTA
GGGCCAAACAAGTTGCACCTGTGACAGTGGCAGCCGACGTGGATGGCAATGCAGCTTACATGAA
TTTGGGGATAGGGCAACGGCCAGACAGCGTTCACAAGATTTTTGGTCAAGATTTGAATTCAATT
CATGCTCAAAGTCGGGAAAAGCTTGTCGTCAGAAAGCCAGACGAAAGAAAGTGGGATGCTTATC
CAAATGGAAACAGAATTACTACGTTTCCAAACACGTGCAATGGTCTGAATAACACCGCGTGGCC
ACAGTTCAATCATGTCAAGCCAGGCAATCCCGTGGAACTTTATAGTCCCCAGACACAACCAAAC
AATTTACAGGGGTTTGTTAATGGAGCAAGTCGTGTAAACAATTTTCAGCATCAAAAGCCAGCTC
AAATGCAAATTGACTTTACAGGAGCGACATCAAGACCTGTAGTTTCAAGACCTGTAAGTGCTGA
GTCTGAGCATTCAGATATTGAGGCTTCATGCAGAGAAGAGAGTGCAGTTCTAGCTGCTGATGAA
AAGAGGCCTCGGAAGCGTGGTCGGAAACCTGCCAATGGAAGAGAAGAGCCCCTCAATCATGTCG
AGGCAGAGAGGCAGCGAAGAGAGAAGCTGAATCAGCGGTTTTACGCACTGAGAGCTGTTGTACC
CAATATTTCTAAAATGGATAAAGCATCCCTTTTGGGAGATGCCATTGCTTACATAACCGAGCTG
CAGAAGAAGCTTAGGGACCTGGAATCAGAAAAGGAGAAAGTGGGAAGCACCTCAAGAGATGCAT
TAGCTTTTGAGAAGAACGCGAGTTCTGGCGTTGAAAATCAAGTGCCTGAAATCGATATTGAAGC
CGCACGTGATGAAGTTATTGTGAAGGTGACCTGTCCATTGGATGCTCATCCTGTATCTGGAGTC
ATTCAAGCATTCAAAGAAACACAGGTGAACGTCTTGGATTCAAAACTGGCCACAGGTAATGAAA
CAGTGTTTCACACATTTGTTGTCAAGTCTCAAGGATCAGACAAGTTAACAAAGGAAAAGTTGAT
AGCAGCAATGTCCCGAGATCGTGAGGCAACCTCACAATCTTCGTGATCATTTTGGTAACGAAAC
CTATGTAATTGACATAGTCATTAGGCATAAATTTTAGTTACAACTCTTTTTATGAAAGCTTCTG
CAGAGATTTTGGTATACAGCAGTTGAAGCTAGGATTAGGAAATATTTATCTCTTGACCTAGTAA
ACATAGCAGTTTTATGTGTTTTGTTTATAGTATACATGTCCTGTCATAGAAGTATTTTTTCTTT

GGAAGATGAATTATGTTGTAGAAAACAACTATCAAGTTTTCTGAACTTCATTTCTAATCACTTC
TTTGAGTTATATTTCTTTTCTTGGTTCTGTTGTTGTCTGAACTCTTACAAAAGTCCAG

Alignment of the retrieved nucleotide sequences retrieved from MPGR and PhytoMetaSyn databases.



Alignment of the retrieved nucleotide sequences retrieved from MPGR and PhytoMetaSyn databases, and the sequenced *CrTF194*.



CrTF246

Tag and MPGR best hit alignment

>cra_locus_5262_iso_1_len_2814_ver_3
Length = 2814

Minus Strand HSPs:

Score = 344 (57.7 bits), Expect = 4.7e-13, Sum P(2) = 4.7e-13
Identities = 76/82 (92%), Positives = 76/82 (92%), Strand = Minus / Plus

```
Query:   108 TTAAGTTCTCAGCAGTGGCAATATTGTCATTACGAGGAATCCACGAAAGGTTCACCGAGC
49
             ||||||||||||||||||||||||||||||| |||||||||| |||||||||| |||
Sbjct:    1785  TTAAGTTCTCAGCAGTGGCAATATTGTCATTACAAGGAATCCAC-AAAGGTTCACC-AGC
1842

Query:    48 ATGAGCCGTTCTAATAAGTTGC 27
             |||||| |||||||||||| |
Sbjct:  1843 TTGAGCCATTCTAATAAGTTCC 1864
```

MPGR - Retrieved protein sequence

```
>cra_locus_5262_iso_1_len_2814_ver_3
MFDSHHHLLDMTHKTPENEMDLIRDDEFESKSGTDIMEAPSGDDQDPNQRPKRKRYHRHTQHQI
QEMESFFKECPHPDDKQRKELGRRLGLEPLQVKFWFQNKRTQMKAQHERHENTQLRNENEKLRA
ENIRYKEALSNATCPNCGGPAAIGEMSYDEQHLRIENARLREEIDRISGIAAKYVGKPMLSYPH
IPGGGGGGSSRSLDLGVGNFTPPAGMVGEIYGAASDLLRSVSGPTEADKPMIIELAVAAMEELI
RMAQAGEPLWIPCNDNIATAENLNEDEYVRTFPRGIGPKPLGLKSEASRESAVVIMNHINLVEI
LMDVNQWSSVFSSIVSRALTLEVLSTGVAGNYNGALQVMTAEFQVPTPLVPTRENYFVRYCKQH
ADGTWAVVDVSLDNLRPTSVSRCRRRPSGCLIQELPNGYSKVTWVEHVEVDDRAVHNIYRGLVN
TGLAFGAKRWVATLDRQCERLASALANNIPAGDIGVITTPEGRKSMLKLAERMVMSFCAGVGAS
TAHTWTTLSGSGADDVRVMTRKSMDDPGRPPGIVLSAATSFWLPVPPKRVFDFLRDENSRSEWD
ILSNGGLVQEMAHIANGRDPGNSVSLLRVNSANSSQSNMLILQESSTDSTGSYVIYAPVDIVAM
NVVLSGGDPDYVALLPSGFAILPDGPPNNGGGVVEVGTGGSLLTVAFQILVDSVPTAKLSLGSV
ATVNSLIKCTVERIKAALACENAXNSTPSIGPQVRLKRMKESKGHRKEVFKLISIGTGGDLKGE
KVKNAPYVILAVWCLQGNXEEERISPRCKGIGSGIDFTKYRQTNQ
```

MPGR- Retrieved nucleotide sequence

```
>cra_locus_5262_iso_1_len_2814_ver_3
TTCAACCAAACATGTTTGATAGTCACCATCACTTACTTGATATGACTCATAAAACACCTGAAAA
TGAGATGGACCTCATTAGGGATGATGAATTTGAGAGCAAATCTGGTACTGATATAATGGAAGCT
CCTTCTGGTGATGATCAAGATCCTAATCAACGTCCCAAAAGAAAGCGTTACCATCGTCACACTC
AGCATCAGATTCAAGAAATGGAATCTTTCTTCAAGGAATGTCCACATCCAGATGACAAACAAAG
GAAAGAACTTGGTCGGCGATTAGGCTTAGAGCCCTTGCAAGTCAAATTTTGGTTTCAGAATAAA
CGTACCCAAATGAAGGCTCAACATGAAAGGCATGAAAACACACAATTAAGGAACGAGAATGAAA
AGCTTCGTGCGGAAAATATTCGATACAAAGAGGCTCTAAGCAATGCCACTTGCCCCAATTGCGG
TGGACCAGCAGCGATTGGGGAGATGTCTTATGATGAACAACATTTAAGAATCGAGAATGCTCGT
TTAAGAGAGGAGATTGATAGGATATCAGGAATAGCTGCGAAATATGTAGGAAAGCCAATGTTAT
CTTATCCACATATTCCAGGAGGAGGAGGAGGAGGATCATCACGTTCCCTAGATCTTGGAGTTGG
GAATTTCACGCCGCCAGCAGGAATGGTCGGAGAAATATATGGGGCGGCTAGTGACCTTCTTAGG
```

TCAGTTTCTGGTCCAACTGAAGCTGATAAGCCGATGATAATTGAATTGGCAGTTGCCGCAATGG
AGGAACTTATTAGAATGGCTCAAGCTGGTGAACCTTTGTGGATTCCTTGTAATGACAATATTGC
CACTGCTGAGAACTTAAATGAAGATGAATATGTTCGGACATTTCCACGTGGAATTGGACCAAAA
CCTTTGGGATTAAAATCTGAAGCTTCTCGAGAATCCGCTGTTGTTATTATGAATCATATTAACC
TCGTTGAAATTCTCATGGATGTGAATCAATGGTCAAGTGTGTTCTCAAGCATTGTTTCAAGAGC
ATTGACCTTGGAGGTGCTGTCAACTGGTGTAGCAGGAAACTATAATGGAGCTTTACAAGTGATG
ACAGCGGAATTTCAGGTTCCCACACCACTGGTTCCAACTCGTGAGAATTACTTTGTGAGATACT
GTAAACAACATGCGGATGGGACATGGGCTGTTGTTGATGTTTCTTTGGATAATCTTCGACCTAC
TTCTGTGTCTAGATGTCGTAGAAGGCCTTCTGGTTGTTTGATTCAAGAATTGCCAAATGGTTAT
TCAAAGGTAACATGGGTGGAACATGTTGAGGTCGATGATAGAGCAGTTCATAACATCTATAGAG
GACTAGTTAATACAGGTCTTGCATTTGGAGCAAAACGCTGGGTTGCTACTTTGGATCGACAATG
TGAACGTCTTGCAAGTGCATTAGCAAATAATATCCCCGCTGGAGATATTGGAGTAATAACGACT
CCAGAAGGTAGAAAAAGTATGTTAAAACTAGCGGAGAGAATGGTAATGAGCTTCTGTGCTGGTG
TGGGAGCTTCAACAGCACATACTTGGACTACATTATCTGGCAGTGGTGCTGATGATGTTCGGGT
CATGACCCGAAAAAGTATGGATGATCCAGGAAGACCTCCTGGCATTGTTTTAAGTGCAGCTACT
TCTTTTTGGCTTCCTGTTCCTCCTAAAAGAGTTTTTGATTTTCTTCGTGATGAGAACTCCAGAA
GTGAGTGGGATATACTATCAAATGGAGGTCTAGTTCAAGAAATGGCACATATTGCCAATGGTCG
TGACCCAGGCAATTCTGTCTCTTTACTCCGTGTTAACAGTGCAAATTCAAGCCAGAGTAACATG
TTAATCTTGCAAGAAAGCAGTACAGATTCAACTGGTTCTTACGTTATTTATGCACCAGTTGATA
TTGTTGCCATGAATGTTGTCTTAAGTGGTGGTGACCCTGACTATGTTGCTTTATTGCCATCTGG
TTTTGCTATCCTGCCTGATGGACCTCCAAATAATGGTGGAGGCGTTGTTGAGGTTGGGACAGGT
GGATCACTTCTAACAGTTGCATTTCAGATATTGGTTGATTCTGTACCAACAGCAAAACTTTCAC
TTGGATCAGTTGCAACTGTTAATAGCCTTATCAAATGTACTGTTGAAAGGATTAAAGCTGCATT
AGCCTGTGAAAATGCTTAG

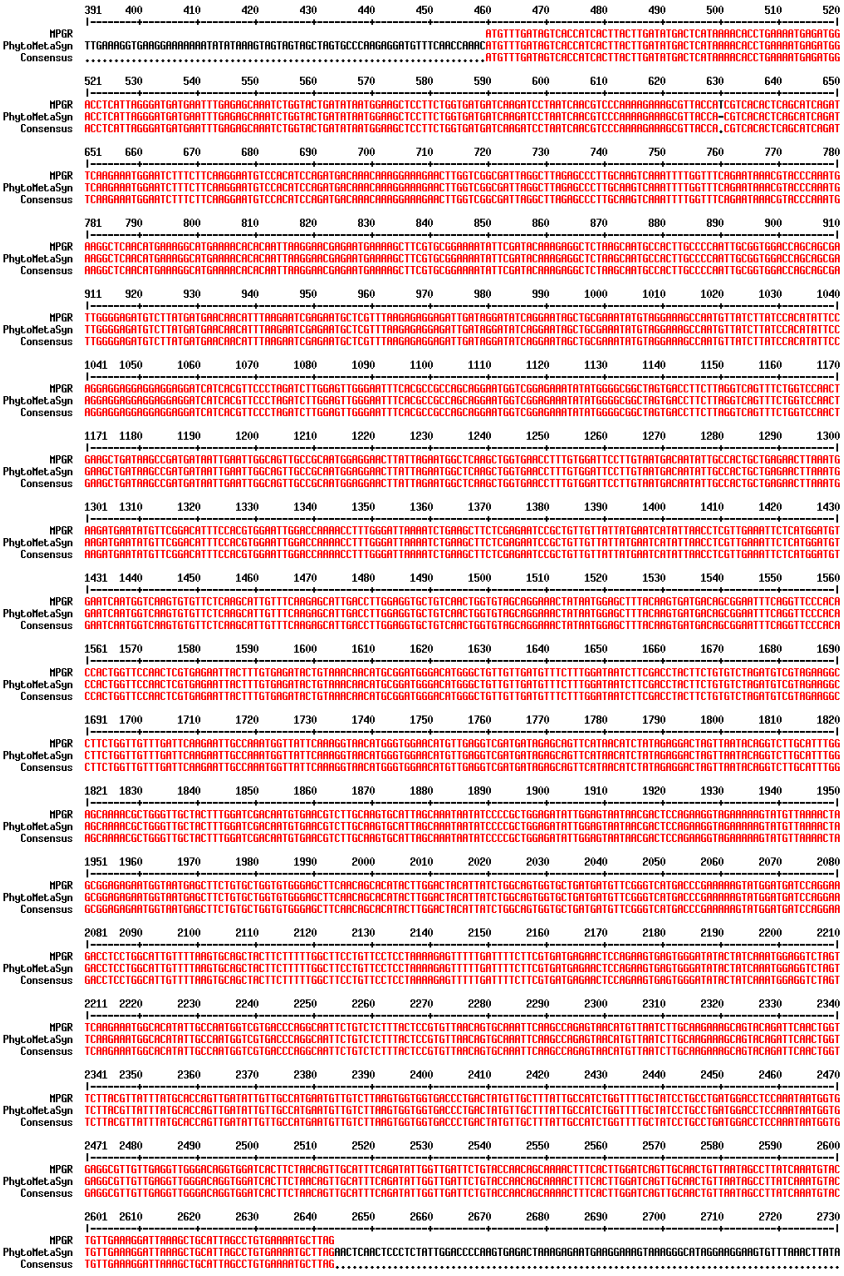
PhytoMetaSyn - Retrieved nucleotide sequence

>gnl|MAGPIE|cro.CROAJ1VD_velvet--Contig2094 No definition line found

CTTCACCCTCTCCCTATCTCTGTTTTCTCTCTCTTAGTCTCATCCTCAGTATTCTTCTTAAATT
CTATTCACAAGAAAGTCATCCATCATTCATAATTCAAACCAGATCTCGCCCTACTATTCTCTGT
TTGGCAAACCCTAAAAAAAACCCCAACAACCTTATTTAGGGTTTTCCTTTCTTTCTTCTCAAAA
ACTTCAAAGTCTCCAAAGAAGCTCCTCCATTCTTATGCTTGAAGTTGTTGATCAAAATCATTAA
GGAATACTTTTGGTTAATTTGAAGTTTTGCGAGTAAAAGAAACAATTAAACATGTCCAAGAGAA
AGCTACAACAGATTATAACTCAAATTAGGATTTTTTCTTCGGAGATCAACTTTTTCAAGATTTT
AAGAAATTGAAAGGTGAAGGAAAAAAATATATAAAGTAGTAGTAGCTAGTGCCCAAGAGGATGT
TTCAACCAAACATGTTTGATAGTCACCATCACTTACTTGATATGACTCATAAAACACCTGAAAA
TGAGATGGACCTCATTAGGGATGATGAATTTGAGAGCAAATCTGGTACTGATATAATGGAAGCT
CCTTCTGGTGATGATCAAGATCCTAATCAACGTCCCAAAAGAAAGCGTTACCACGTCACACTCA
GCATCAGATTCAAGAAATGGAATCTTTCTTCAAGGAATGTCCACATCCAGATGACAAACAAAGG
AAAGAACTTGGTCGGCGATTAGGCTTAGAGCCCTTGCAAGTCAAATTTTGGTTTCAGAATAAAC
GTACCCAAATGAAGGCTCAACATGAAAGGCATGAAAACACACAATTAAGGAACGAGAATGAAAA
GCTTCGTGCGGAAAATATTCGATACAAAGAGGCTCTAAGCAATGCCACTTGCCCCAATTGCGGT
GGACCAGCAGCGATTGGGGAGATGTCTTATGATGAACAACATTTAAGAATCGAGAATGCTCGTT
TAAGAGAGGAGATTGATAGGATATCAGGAATAGCTGCGAAATATGTAGGAAAGCCAATGTTATC
TTATCCACATATTCCAGGAGGAGGAGGAGGAGGATCATCACGTTCCCTAGATCTTGGAGTTGGG
AATTTCACGCCGCCAGCAGGAATGGTCGGAGAAATATATGGGGCGGCTAGTGACCTTCTTAGGT
CAGTTTCTGGTCCAACTGAAGCTGATAAGCCGATGATAATTGAATTGGCAGTTGCCGCAATGGA

GGAACTTATTAGAATGGCTCAAGCTGGTGAACCTTTGTGGATTCCTTGTAATGACAATATTGCC
ACTGCTGAGAACTTAAATGAAGATGAATATGTTCGGACATTTCCACGTGGAATTGGACCAAAAC
CTTTGGGATTAAAATCTGAAGCTTCTCGAGAATCCGCTGTTGTTATTATGAATCATATTAACCT
CGTTGAAATTCTCATGGATGTGAATCAATGGTCAAGTGTGTTCTCAAGCATTGTTTCAAGAGCA
TTGACCTTGGAGGTGCTGTCAACTGGTGTAGCAGGAAACTATAATGGAGCTTTACAAGTGATGA
CAGCGGAATTTCAGGTTCCCACACCACTGGTTCCAACTCGTGAGAATTACTTTGTGAGATACTG
TAAACAACATGCGGATGGGACATGGGCTGTTGTTGATGTTTCTTTGGATAATCTTCGACCTACT
TCTGTGTCTAGATGTCGTAGAAGGCCTTCTGGTTGTTTGATTCAAGAATTGCCAAATGGTTATT
CAAAGGTAACATGGGTGGAACATGTTGAGGTCGATGATAGAGCAGTTCATAACATCTATAGAGG
ACTAGTTAATACAGGTCTTGCATTTGGAGCAAAACGCTGGGTTGCTACTTTGGATCGACAATGT
GAACGTCTTGCAAGTGCATTAGCAAATAATATCCCCGCTGGAGATATTGGAGTAATAACGACTC
CAGAAGGTAGAAAAAGTATGTTAAAACTAGCGGAGAGAATGGTAATGAGCTTCTGTGCTGGTGT
GGGAGCTTCAACAGCACATACTTGGACTACATTATCTGGCAGTGGTGCTGATGATGTTCGGGTC
ATGACCCGAAAAAGTATGGATGATCCAGGAAGACCTCCTGGCATTGTTTTAAGTGCAGCTACTT
CTTTTTGGCTTCCTGTTCCTCCTAAAAGAGTTTTTGATTTTCTTCGTGATGAGAACTCCAGAAG
TGAGTGGGATATACTATCAAATGGAGGTCTAGTTCAAGAAATGGCACATATTGCCAATGGTCGT
GACCCAGGCAATTCTGTCTCTTTACTCCGTGTTAACAGTGCAAATTCAAGCCAGAGTAACATGT
TAATCTTGCAAGAAAGCAGTACAGATTCAACTGGTTCTTACGTTATTTATGCACCAGTTGATAT
TGTTGCCATGAATGTTGTCTTAAGTGGTGGTGACCCTGACTATGTTGCTTTATTGCCATCTGGT
TTTGCTATCCTGCCTGATGGACCTCCAAATAATGGTGGAGGCGTTGTTGAGGTTGGGACAGGTG
GATCACTTCTAACAGTTGCATTTCAGATATTGGTTGATTCTGTACCAACAGCAAAACTTTCACT
TGGATCAGTTGCAACTGTTAATAGCCTTATCAAATGTACTGTTGAAAGGATTAAAGCTGCATTA
GCCTGTGAAAATGCTTAGAACTCAACTCCCTCTATTGGACCCCAAGTGAGACTAAAGAGAATGA
AGGAAAGTAAAGGGCATAGGAAGGAAGTGTTTAAACTTATAAGCATAGGAACAGGAGGAGACTT
AAAAGGGGAAAAAGTCAAGAACGCACCATATGTAATCCTTGCCGTGTGGTGTTTGCAAGGAAAT
TAAGAAGAAGAACGCATTTCACCACGCTGCAAGGGTATAGGTTCGGGTATTGACTTCACCAAAT
ATCGTCAAACCAATCAATGAATGATTTTCAAGTAAAAACTTAATTATATAACGCTTCCTCAAAT
TTCACGAAGGCATGTCATGTATTTTGCTTAATTATTAGGAATTAATTAGGCGTTTCGATTCATC
AATATGTATTATGTGTATTGAAAGTTATTTAGATATATTAACTTCTTTGCTCGAGTCTAGTTTA
CTATATATTGCTAACAAACTATGGTTTCCTTTTCGAACTAAACTCTCTATGTCTTGATTTTAAC
CTCCTATCAATTTGTTTATTGGCAATTGCTGAGGAAGGTTTATTTTCTAATGAAAGACCTTTTT
ATATACCGCCTAGGTTTTTCC

Alignment of the retrieved nucleotide sequences from MPGR and PhytoMetaSyn databases



7.7 Appendix 7. Results of the BLAST search for the retrieval of the full coding sequences of the CrTF61 gene.

Tag and MPGR best hit alignment

cra_locus_45433_iso_1_len_858_ver_3
Length = 858

Plus Strand HSPs:

Score = 2286 (349.0 bits), Expect = 2.8e-98, P = 2.8e-98
Identities = 457/458 (99%), Positives = 458/458 (100%), Strand = Plus / Plus

```
Query:                                                                   1
GATCCGAGCCAATAACAACACCAAAAAGAATCCCAGATGCGCCGCCGATGGAAACGAGCT 60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:                                                                 336
GATCCGAGCCAATAACAACACCAAAAAGAATCCCAGATGCGCCGCCGATGGAAACGAGCT 395

Query:                                                                  61
TCTGAGATTCCATTGTACGACGCTCACGTGTTCGCTCGGCGCTCGCGGCTCGTCCAGCTT 120
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   396 TCTGAGATTCCATTGTACGACGCTCACGTGTTCGCTCGGCGCTCGCGGCTCGTCCAGCTT
455

Query:                                                                 121
GTGCGGCTCCGTACCGGGCTGTGGAGTTTGCACTATCATAAGGCACGGCTTTCAAGGAAA 180
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:                                                                 456
GTGCGGCTCCGTACCGGGCTGTGGAGTTTGCACTATCATAAGGCACGGCTTTCAAGGAAA 515

Query:                                                                 181
CAAGTCTAGTGGGGTCCGCACTACCGCTAGCAGTGGTAGGGCCCACGACTGCCTAGGTAT 240
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:                                                                 516
CAAGTCTAGTGGGGTCCGCACTACCGCTAGCAGTGGTAGGGCCCACGACTGCCTAGGTAT 575

Query:                                                                 241
CCCCGCCGCCGAAGCCGAGGGGGGTCGTAGTAGTAGAGCTATGTTAGTGTGTCGTGTGAT 300
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:                                                                 576
CCCCGCCGCCGAAGCCGAGGGGGGTCGTAGTAGTAGAGCTATGTTAGTGTGTCGTGTGAT 635

Query:                                                                 301
TGCTGGAAGAGTGAAGCGCGTGGTGGATGATACGGGTGCGGAAGCGGCTCCTGAGGAGGA 360
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:                                                                 636
TGCTGGAAGAGTGAAGCGCGTGGTGGATGATACGGGTGCGGAAGCGGCTCCTGAGGAGGA 695

Query:                                                                 361
TGCTGGGGCGTCGTTGGCGGCTCCTCTTGGCTCGTATGACTCGTTAGCCGGATACGCCGG 420
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:                                                                 696
TGCTGGGGCGTCGTTGGCGGCTCCTCTTGGCTCGTATGACTCGTTAGCCGGATACGCCGG 755

Query:   421 GATTTATTCGAATCTTGAGGAACTTTACGTGTTTAATS 458
              |||||||||||||||||||||||||||||||||||+
Sbjct:   756 GATTTATTCGAATCTTGAGGAACTTTACGTGTTTAATC 793
```

MPGR- Retrieved nucleotide sequence

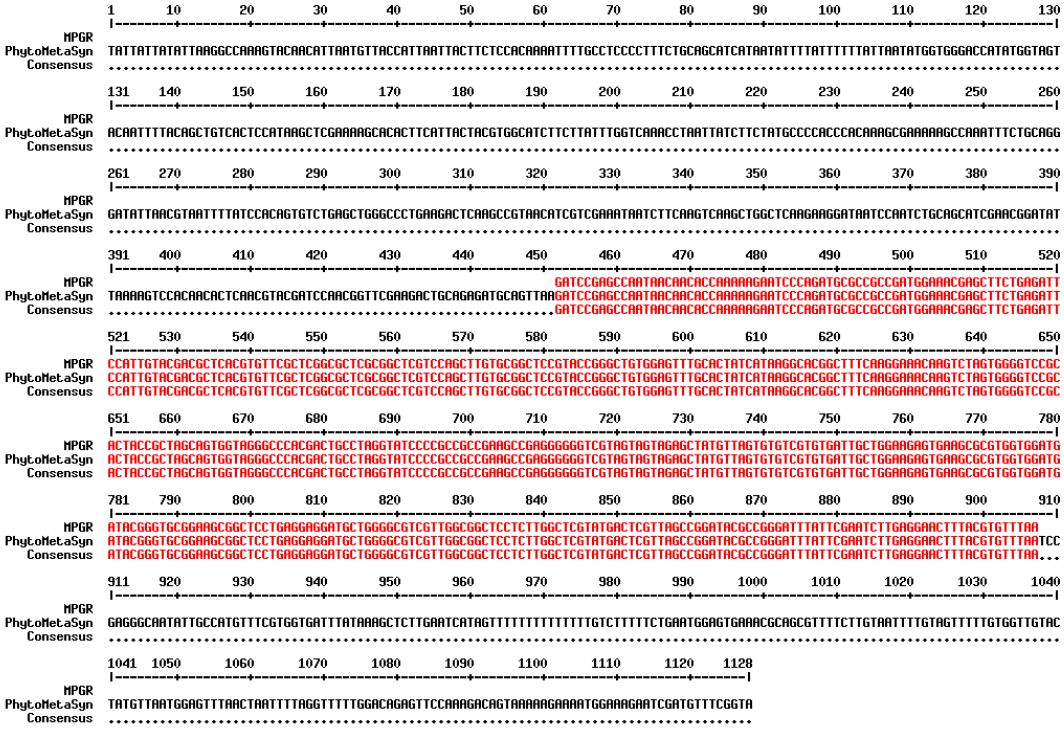
GATCCGAGCCAATAACAACACCAAAAAGAATCCCAGATGCGCCGCCGATGGAAACGAGCTTCTG
AGATTCCATTGTACGACGCTCACGTGTTCGCTCGGCGCTCGCGGCTCGTCCAGCTTGTGCGGCT
CCGTACCGGGCTGTGGAGTTTGCACTATCATAAGGCACGGCTTTCAAGGAAACAAGTCTAGTGG
GGTCCGCACTACCGCTAGCAGTGGTAGGGCCCACGACTGCCTAGGTATCCCCGCCGCCGAAGCC
GAGGGGGGTCGTAGTAGTAGAGCTATGTTAGTGTGTCGTGTGATTGCTGGAAGAGTGAAGCGCG
TGGTGGATGATACGGGTGCGGAAGCGGCTCCTGAGGAGGATGCTGGGGCGTCGTTGGCGGCTCC
TCTTGGCTCGTATGACTCGTTAGCCGGATACGCCGGGATTTATTCGAATCTTGAGGAACTTTAC
GTGTTTAA

PhytoMetaSyn - Retrieved nucleotide sequence

>gnl|MAGPIE|cro.CROWL1VD_velvet--Singlet22282 No definition line found

TATTATTATATTAAGGCCAAAGTACAACATTAATGTTACCATTAATTACTTCTCCACAAAATTT
TGCCTCCCCTTTCTGCAGCATCATAATATTTTATTTTTTATTAATATGGTGGGACCATATGGTA
GTACAATTTTACAGCTGTCACTCCATAAGCTCGAAAAGCACACTTCATTACTACGTGGCATCTT
CTTATTTGGTCAAACCTAATTATCTTCTATGCCCCACCCACAAAGCGAAAAAGCCAAATTTCTG
CAGGGATATTAACGTAATTTTATCCACAGTGTCTGAGCTGGGCCCTGAAGACTCAAGCCGTAAC
ATCGTCGAAATAATCTTCAAGTCAAGCTGGCTCAAGAAGGATAATCCAATCTGCAGCATCGAAC
GGATATTAAAAGTCCACAACACTCAACGTACGATCCAACGGTTCGAAGACTGCAGAGATGCAGT
TAAGATCCGAGCCAATAACAACACCAAAAAGAATCCCAGATGCGCCGCCGATGGAAACGAGCTT
CTGAGATTCCATTGTACGACGCTCACGTGTTCGCTCGGCGCTCGCGGCTCGTCCAGCTTGTGCG
GCTCCGTACCGGGCTGTGGAGTTTGCACTATCATAAGGCACGGCTTTCAAGGAAACAAGTCTAG
TGGGGTCCGCACTACCGCTAGCAGTGGTAGGGCCCACGACTGCCTAGGTATCCCCGCCGCCGAA
GCCGAGGGGGGTCGTAGTAGTAGAGCTATGTTAGTGTGTCGTGTGATTGCTGGAAGAGTGAAGC
GCGTGGTGGATGATACGGGTGCGGAAGCGGCTCCTGAGGAGGATGCTGGGGCGTCGTTGGCGGC
TCCTCTTGGCTCGTATGACTCGTTAGCCGGATACGCCGGGATTTATTCGAATCTTGAGGAACTT
TACGTGTTTAATCCGAGGGCAATATTGCCATGTTTCGTGGTGATTTATAAAGCTCTTGAATCAT
AGTTTTTTTTTTTTTTGTCTTTTTCTGAATGGAGTGAAACGCAGCGTTTTCTTGTAATTTTGTA
GTTTTTGTGGTTGTACTATGTTAATGGAGTTTAACTAATTTTAGGTTTTTGGACAGAGTTCCAA
AGACAGTAAAAAGAAAATGGAAAGAATCGATGTTTCGGTA

Alignment of the retrieved nucleotide sequences from MPGR and PhytoMetaSyn databases



Protein sequence translated from the nucleotide sequence retrieved from PhytoMetaSyn database

MLPLITSPQNFASPFCSIIIFYFLLIWWDHMVVQFYSCHSISSKSTLHYYVASSYLVKPNYLLC
PTHKAKKPNFCRDINVILSTVSELGPEDSSRNIVEIIFKSSWLKKDNPICSIERILKVHNTQRT
IQRFEDCRDAVKIRANNNTKKNPRCAADGNELLRFHCTTLTCSLGARGSSSLCGSVPGCGVCTI
IRHGFQGNKSSGVRTTASSGRAHDCLGIPAAEAEGGRSSRAMLVCRVIAGRVKRVVDDTGAEAA
PEEDAGASLAAPLGSYDSLAGYAGIYSNLEELYVFNPRAILPCFVVIYKALES-